

МЕТОД ЗА ОЦЕНКА И ИЗБОР НА РЕФЕРЕНТНА АРХИТЕКТУРА ЗА ИНТЕРНЕТ НА ОБЕКТИТЕ

Автореферат на дисертационен труд за
придобиване на образователна и научна
степен "доктор"

инж. Явор Чавдаров Христов

Научен ръководител: доц. д-р Любен Боянов

НАУЧНО НАПРАВЛЕНИЕ "ПРИЛОЖЕНИЕ НА ИТ В ИКОНОМИКАТА"

КАТЕДРА "ИНФОРМАЦИОННИ ТЕХНОЛОГИИ И КОМУНИКАЦИИ"

ФАКУЛТЕТ "ПРИЛОЖНА ИНФОРМАТИКА И СТАТИСТИКА"

УНИВЕРСИТЕТ ЗА НАЦИОНАЛНО И СВЕТОВНО СТОПАНСТВО

СОФИЯ 2021 Г.

Дисертационният труд е обсъден и допуснат до публична защита на редовно заседание на катедрен съвет на катедра "Информационни Технологии и Коммуникации" на Университета за Национално и Световно Стопанство - София (УНСС), включващо членове на катедрата, състояло се на 12 март 2021 г. Авторът на дисертационния труд е задочен докторант в катедрата.

Публичната защита на дисертационния труд е насрочена за 01 юни 2021 г., от 13:00 ч. в зала "Тържествена" на УНСС, централно фоайе (№ 2032А), пред научно жури в състав:

Външни членове:

1. проф. д-р Красен Стефанов;
2. проф. д-р Юлиан Василев;
3. проф. д-р Красимир Шишманов.

Вътрешни членове:

1. доц. д-р Моника Цанева;
2. доц. д-р Любен Боянов.

Материалите по защитата са публикувани на Интернет страницата на Университета за Национално и Световно Стопанство - София (<https://www.unwe.bg>) и са на разположение на интересуващите се в сектор "Научни съвети и конкурси" към отдел "Наука" на УНСС.

СЪДЪРЖАНИЕ

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД	4
1. АКТУАЛНОСТ	4
2. ОБЕКТ И ПРЕДМЕТ НА ДИСЕРТАЦИОННИЯ ТРУД	5
3. РАБОТНИ ХИПОТЕЗИ НА ДИСЕРТАЦИОННИЯ ТРУД.....	6
4. ЦЕЛИ И ЗАДАЧИ НА ДИСЕРТАЦИОННИЯ ТРУД.....	6
5. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД	7
6. ОБЕМ И СТРУКТУРА НА ДИСЕРТАЦИОННИЯ ТРУД	8
КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД	10
1. ВЪВЕДЕНИЕ.....	10
2. ГЛАВА I. ИНТЕРНЕТ НА ОБЕКТИТЕ, ГОЛЕМИ ДАННИ И РЕФЕРЕНТНИ АРХИТЕКТУРИ.....	10
3. ГЛАВА II. СРЕДСТВА ЗА ОБРАБОТКА НА ГОЛЕМИ ДАННИ И ОПИСАНИЕ НА КОМПОНЕНТИТЕ	21
4. ГЛАВА III. МЕТОД ЗА ОЦЕНКА НА РЕФЕРЕНТНИ АРХИТЕКТУРИ. РАЗРАБОТВАНЕ НА МОДЕЛИ ЗА ИНО РА.	33
5. ГЛАВА IV. ПРИЛОЖЕНИЕ НА СЪЗДАДЕНАТА РА ЗА ОБРАБОТКА И АНАЛИЗ НА РАЗЛИЧНИ ТИПОВЕ ДАННИ.....	46
6. ЗАКЛЮЧЕНИЕ	62
7. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД.....	63
ЛИТЕРАТУРА.....	64
СПИСЪК НА ФИГУРИТЕ	65
СПИСЪК НА ТАБЛИЦИТЕ	66

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. АКТУАЛНОСТ

Днес Интернет на Обектите (Ино), известен и с английския термин Internet of Things (IoT) е навсякъде – сензори измерват физически стойности като температура, влажност, налягане, цифрови идентификатори са прикрепени към стоки, транспортни средства даже и живи същества, умни устройства свързват и управляват процеси и уреди в умни офиси и умни домове. Всички подобни устройства генерират огромни количества данни, като цифровите данни вече не са само в домейна на компютрите и Интернет, нито пък са само в мобилните телефони, а се ползват и намират приложение във всяка сфера на човешката дейност. Тези данни се използват от информационни системи за отчитане, съхраняване и обратна връзка към обектите, от които са произлезли тези данни. Създаването и функционирането на тези информационни системи е по-различно от традиционните системи, които работят с данни, генерирани от компютри, таблети, мобилни устройства и различни сървъри в Интернет. Различни са изискванията към такива системи, различна е архитектурата на тези системи, както и протоколите за обмен на данни. Заедно с това, наличието на много и разнообразна информация предоставя нови, качествено различни възможности на хората да следят и управляват света около себе си. Това показва, че работата с големи данни и Ино е не само актуална тема, а и такава, която ще остане в ползрението на хората и сферата на информационните технологии (ИТ) още десетилетия наред. Държавни институции, големи корпорации, средни и малки фирми, стандартизиращи организации и много други ще работят в насока за по-широко разпространение и приложение на системи базирани на Ино и последващата обработка, съхраняване и използване на големите данни в тези системи.

Редица анализатори считат, че светът се намира на прага на нова, революционна промяна, в основата на която са информационните и комуникационни технологии. Има оценки, че очаквания икономически ефект от развитието на ИТ и по-специално на Ино може да достигне през 2025 г. между 4 и 11 млрд. USD на година [1]. Тези оценки са направени на базата на прогнозите на световната банка за брутен вътрешен продукт (БВП) от 99.5 млрд. USD през 2025 г. и в основата им лежат намаляването на цените на комуникациите, цените за производство на хардуер, увеличаване на скоростите и честотните ленти, напредъкът в изчислителните процеси, все по-важното значение на облачните технологии и, разбира се,

увеличаването на броя на свързаните посредством Интернет Ино устройства. Според горесцитирания източник [1], през 2025 г. се очаква Ино да допринасят за обем между 4 и 11 процента от световния БВП.

Ино е обект на изследователски и научно-приложен интерес както от научната и академична общност (data science), така и от редица търговски, производствени, транспортни и много други компании (data analytics и data engineering).

Някои от основните предимства на Ино са:

- достъп до информацията от всяко устройство по всяко време и от всяко място;
- подобрена комуникация между различните Ино устройства;
- свързването и трансфера на данни от Ино устройствата спестява време и средства;
- възможност за автоматизация на различни организационни и бизнес процеси;
- подобряване на качеството и минимизиране на човешките грешки.

Сред недостатъците на Ино можем да споменем:

- потенциалните проблеми свързани с неоторизиран достъп до информацията;
- огромните количества данни изправят всички потребители на Ино пред редица предизвикателства за събирането, съхранението и управлението на тези Ино данни;
- потенциални софтуерни проблеми (бъгове) могат да бъдат бързо разпространени в огромен брой Ино устройства;
- липсата на утвърдени международни стандарти за Ино устройства е предизвикателство за интеграцията на Ино устройства от различни производители.

2. ОБЕКТ И ПРЕДМЕТ НА ДИСЕРТАЦИОННИЯ ТРУД

Обект на настоящия дисертационен труд са референтните архитектури (РА) за Интернет на обектите, където се задават основни понятия и блокове за изграждане на системи, приемащи, съхраняващи и обработващи данни от Интернет на обектите.

Предмет на дисертационния труд е методологията за създаване на референтна архитектура за големи данни, начините за оценка на тази архитектура и методите за интеграция на отделните компоненти.

Изследователският въпрос, поставен в настоящия дисертационен труд е "Как се създава референтна архитектура за обработка на данни от Интернет на обектите, по какъв начин може да се получи количествена оценка за нея и какви са подходите за ефективно използване на функционалните компоненти на референтната архитектура?".

3. РАБОТНИ ХИПОТЕЗИ НА ДИСЕРТАЦИОННИЯ ТРУД

Основните работни хипотези за настоящия дисертационен труд са:

Хипотеза I:

Възможно е да се създаде референтна архитектура, приложима за обработка, съхранение и работа с данни от Интернет на обектите.

Хипотеза II:

Възможно е да се даде количествена оценка за референтна архитектура, като се създаде методология за оценка по различни показатели.

Хипотеза III:

Възможно е да се постигне ефективна интеграция на компонентите на референтната архитектура, за работа с големи данни.

4. ЦЕЛИ И ЗАДАЧИ НА ДИСЕРТАЦИОННИЯ ТРУД

Основните цели на представения дисертационен труд са дефиниране на метод за количествена оценка на Ино РА и създаване на максимално гъвкава и оперативно съвместима Ино РА, която да бъде приложена и тествана с реални данни от различни области на Ино.

Задачите, които се изпълняват в процеса на реализация на поставената цел на дисертационния труд са:

- **Изследване и анализ на съвременни референтни архитектури на Интернет на обектите**
 - Обзор и анализ на актуалните проблеми в областта на големите данни и Ино референтни архитектури

- Анализ на използваните средства за обработка на големи данни и на Ино референтни архитектури
- **Изследване и анализ на походите за създаване на Ино референтни архитектури**
 - Идентификация на подходящи компоненти за свързване на Ино референтни архитектури с разнородни Ино източници и извличане и почистване на разнородни Ино данни от тях
 - Идентификация на подходящи компоненти на Ино референтни архитектури за обработка и съхранение на Ино данни
 - Идентификация на подходящи компоненти на Ино за трансформация и презентация на Ино данни
- **Разработване на модел на Ино референтни архитектури**
 - Дефиниране на критерии за оценка на моделите на Ино референтни архитектури
 - Разработване на модел за Ино референтни архитектури
 - Тестване на модела за Ино референтни архитектури с разнородни Ино данни

Представеното изследване може да помогне на специалисти, работещи по теми, свързани с Интернет на обектите и големи данни да използват предложената референтна архитектура и методи, както и използваните конкретни платформи за работа. Ясно е, че всеки проблем или кръг от задачи може да изисква допълнителни функционалности и съответно – разширени архитектурни приложения методи. Заедно с това, рамката и методите за оценка и избор на референтна архитектура могат да помогнат на разработчиците да поставят още в началото критериите, по които да се работи за намиране на оптимална архитектура, която да задоволява очакванията на потребителите.

5. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

Научни приноси

- Създадена е класификация на класовете референтни архитектури, която е резултат от обзор и анализ на най-широко разпространените сред тях, показващ, че не съществува еталонна и общоприета референтна архитектура.

- Създаден е метод за количествена оценка на референтни архитектури, позволяващ оптимизация на процесите на избор на референтна архитектура.
- Проектирана е уникална софтуерна архитектура позволяваща класификация на входящите данни с цел избирането на подходящ специфичен инструмент за анализ.
- Проектираната софтуерна архитектура предоставя комплексно решение за работа с големи данни от постъпването им до средствата за визуализация и използване на бизнес интелигентност.
-

Научно-приложни приноси

- Създадена е работеща платформа за големи данни на базата на референтна архитектура за Ино. Създадената платформа беше тествана успешно в реални условия с данни от различни източници.
- За създадената платформа е характерно свойството взаимозаменяемост на компонентите ѝ. Това позволява оптимизацията нейната производителност и бързодействие - критични параметри на системите за обработка на големи данни.
- Бяха използвани данни от метеорологични сензори, данни от транспортни средства данни и финансови данни. Създадената платформа за Ино може да бъде използвана и за други Ино източници. Беше демонстрирано, че създадената платформа, на базата на предложената референтна архитектура, има широко приложение в различни икономически, логистични, екологични, производствени и други области.

6. ОБЕМ И СТРУКТУРА НА ДИСЕРТАЦИОННИЯ ТРУД

Дисертационният труд е разработен в общ обем от 166 страници, от които 151 съдържат същинското изследване, а 15 са приложения. Структуриран е във Въведение, 4 глави, Заключение, Списък на научните и научно-приложните приноси, Списък на направените публикации по темата на дисертационния труд, Бъдеща работа, Литература, Списъци на фигурите и на таблиците и Списък на термините и съкращенията.

Въведението представя актуалността на проблема, предмета и обекта на изследването. Формулирани са основните хипотези на дисертационния труд, а като негова поставена цел е определено "дефинирането на метод за количествена оценка на Ино РА и създаване на максимално гъвкава и оперативно съвместима Ино РА, която да бъде приложена и тествана с реални данни от различни области

на Ино". Целта на разработката е декомпозирана на задачи, които трябва да бъдат решени последователно.

Първа глава има обзореен характер. Въвеждат се основните понятия използвани в труда и се прави обстоен анализ на редица съществуващи референтни архитектури за Ино. Изводът, че не съществува еталонна и общо приета референтна архитектура дава основание за същинската част на разработката.

Втора глава разглежда основните, спомагателните и общите компоненти използвани за обработка на данни и дефинира изискванията, на които те трябва да отговарят. На тази основа са изведени ключовите компоненти за изграждане на една референтна архитектура за Ино и е мотивиран изборът на тези, които ще бъдат използвани при проектирането на модела за референтна архитектура за Ино. Тази част от дисертационния труд поставя научните основи на изследването и дефинира параметрите на практическата разработка.

Трета глава представя разработения метод за количествена оценка на референтни архитектури за Ино, като неговата приложимост е апробирана за конкретни популярни референтни архитектури. Направен е анализ на възможните начини за създаване на референтни архитектури за Ино. Създадена е гъвкава компонентна референтна архитектура за Ино, на база на мотивиран избор на компонентния метод за изграждането ѝ. Тази глава съдържа основните научни приноси на дисертационния труд.

Четвърта глава представя експерименталната част на дисертационния труд. Създадена е платформа за големи данни, в която са реализирани 3 приложения за приемане, ръчна и автоматична обработка на данни от метеорологични сензори, от транспортни средства и финансови показатели. Прототипите демонстрират широката приложимост както на създадената платформа, така и на стоящата в основата ѝ референтна архитектура. Тази част на дисертационния труд съдържа съществени научно-приложни приноси.

В Заключението се обобщава, че като резултат от разработките, формулираните работни хипотези са доказани и се изразява намерение да се продължат изследванията и разработките в областта на машинното обучение и създаването на предсказващи модели на базата на натрупаните от референтните архитектури големи Ино данни.

КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

1. ВЪВЕДЕНИЕ

Въведението представя актуалността на проблема, предмета и обекта на изследването. Формулирани са основните хипотези на дисертационния труд, а като негова поставена цел е определено "дефинирането на метод за количествена оценка на ИнО РА и създаване на максимално гъвкава и оперативно-съвместима ИнО РА, която да бъде приложена и тествана с реални данни от различни области на ИнО". Целта на разработката е декомпозирана на задачи, които трябва да бъдат решени последователно.

2. ГЛАВА I. ИНТЕРНЕТ НА ОБЕКТИТЕ, ГОЛЕМИ ДАННИ И РЕФЕРЕНТНИ АРХИТЕКТУРИ

Интернет на обектите

Изразът Интернет на обектите (някои автори използват и Интернет на нещата) обозначава огромна мрежа от милиарди обекти, свързани чрез цифрови технологии. Самите обекти може да са предмети, компютри, машини, хора, животни и т.н., които могат да генерират и предават данни, а също така някои от тях – и да приемат и обработват данни. Тъй като този вид комуникация от всеки към всеки е характерен за навлязлата навсякъде световна мрежа Интернет, свързването на (евентуално) всеки обект с други такива по подобен начин се нарича Интернет на обектите. Възможностите и перспективите на ИнО са ясни, поради което развитието на тази област в ИТ сферата стана приоритет от поне десетилетие за всички големи фирми и организации. Бизнесът също обърна внимание на този феномен, като в днешни дни сме свидетели на вече развиващи се и даже утвърдени продукти като умни телефони, умни часовници, умни телевизори. Все по-често чуваме за и виждаме проекти за умни градове, умни къщи и за всякакви други "умни" дейности. Основната идея е, че всеки обект може уникално да бъде дигитално идентифициран, след което да може да предава и приема данни. Това може да стане с жични или безжични мрежи, с което хората (или дигитални устройства и системи) да следят, анализират и даже променят тяхното състояние. На базата на възможностите за подобна дигитализация се разглеждат и невъобразимо големите перспективи на ИнО като средство за отчитане, следене, анализ и въздействие на реалната, заобикаляща хората среда. Автор на словосъчетанието Internet of Things е Кевин Аштън, съосновател на Ауто-идентификаторен център към MIT. Изразът е

използван при представяне на радиочестотни идентификатори (RFID) по време на презентация за фирмата Procter & Gamble (P&G) през 1999 г. [2]. Въпреки появата му в края на 20 век, терминът ИнО става популярен едва след първото десетилетие на следващото столетие. Според "бялата статия" (white paper) на една от най-известните компании в областта на комуникационните технологии Cisco от 2011 г. [3], терминът Internet of Things е "роден" в периода между 2008 и 2009 г., когато броят на свързаните в Интернет устройства е надвишил броя на хората населяващи Земята.

ИнО позволява не само хората да имат по-удобна среда за живот в своите умни домове и градове, но също така тяхната работа и дейност да са по-автоматизирани и с намален разход на труд и място за грешки. С подобряване на ефективността се намалят разходите и отпадъците, подобряват се доставките, намаля се цената на производство и се ускорява процесът на достигане на услугите до крайните потребители, като същевременно последните имат поглед върху пътя на поръчания от тях продукт от фабриката до дома/мястото на получаване. С тези нови функционалности ИнО навлиза в нашето всекидневие и ще продължава да набира скорост, защото все повече бизнес сфери не само наблюдават, но и реализират потенциала на свързване на обектите и извличане на бизнес полза от тази нова световна мрежа.

Функционални архитектурни компоненти на Интернет на обектите

Как може да се разгледат основните съставни блокове на Интернет на обектите от функционална гледна точка? Най-общо може да се посочат следните компонентни блокове:

- крайни устройства като сензори, микроконтролери, RFID и т.н.;
- средства за предаване на данни от крайните устройства – сензорни мрежи, мобилни мрежи, Интернет, гейтуеи, рутери, хъбове и т.н.;
- системи за получаване, съхранение, обработка, визуализация и анализ на данни – смартфони, облачни системи, отделни ИТ системи, аналитични бизнес приложения и т.н.

Основните компоненти на ИнО и тяхната най-обща взаимосвързаност са представени на Фигура 1.

Примерна Ино система



Фигура 1. Структура и връзки на основните компоненти на Ино.

От крайните компоненти може да посочат сензорите, RFID, NFC, микроконтролерите и някои механични устройства.

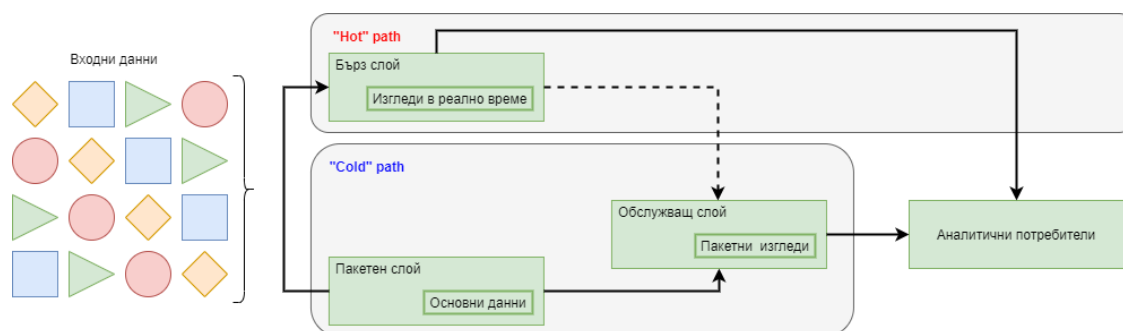
Като средства за предаване на данни от крайните компоненти може да се посочат сензорните мрежи, мобилните мрежи, безжичните мрежи, Интернет, рутери, хъбове и т.н.

Като системи за получаване, съхранение, обработка, визуализация и анализ на данните може да се посочат смартфоните, облачни системи, отделни ИТ системи, аналитични бизнес приложения и т.н.

Тясно свързани с Ино термини са обработка в края (edge computing) и в мъглата (fog computing). Както стана ясно по-горе, сензори и устройства разположени в различни среди в Ино събират и генерират данни от обекти, след което ги предават към облачни среди (публични, частни или хибридни облаци) през специализирани или стандартни гейтуеи (шлюзове). В този процес има още един аспект – локалната обработка и съхранение на данни. Прихващането на повечето данни и предаването им към отдалечен облак не е оптимално за редица приложения. Оттук произтича и идеята изчисления да се изместват към края и/или в мъглата и тази идея се развива все повече през последните години.

Ламбда архитектура за обработка на данни

Ламбда архитектурата (Lambda Architecture - LA) е термин за обща, разширяема и устойчива на грешки архитектура за обработка на данни [4]. LA цели създаване на устойчива система, в която се избягват грешки (както хардуерни, така човешки). Идеята е тази архитектура да обслужва големи потоци данни и натоварвания в различни и многообразни случаи. С използването на тази система се очаква да се постигне ниско закъснение при четене и обновяване. Друг важен момент е, че архитектурата трябва да е линейно скалируема. Потокът на данни се разделя в две посоки – поток за пакетна обработка (batch layer - cold path), който се прилага за по-големи обеми данни и поток за бърза обработка (speed layer - hot path), който се прилага за обработка на данни в реално или почти реално време - Фигура 2.



Фигура 2. Ламбда архитектура.

ИнО предоставя не само възможности, но и проблеми. Сред най-сериозните такива са сигурността на създаваните и обменяни данни, които могат да носят чувствителна и/или конфиденциална информация. Да се предотврати неоторизиран достъп до данните или визуалната информация за нашия дом е от най-голяма важност за всеки индивид. Поверителността на производствените процеси и веригите за доставки са ключови за живота и оцеляване на бизнеса.

Видове данни и големи данни

Данните в съвременния свят обикновено представляват поредица от символи (най-често от цифри с различно представяне), които се интерпретират по определен начин. Данните са формална регистрация на факти в цифров вид, в резултат на дейности като измерване, експерименти, аритметични и логически

операции и др. Те могат да се представят чрез числова стойност, графика, текст, изображение, звук, видео и други. За да се говори за информация трябва данните да бъдат анализирани, интерпретирани и да бъдат извлечени съответните изводи и обяснения на данните. Всички съвременни компютри работят с цифрово представяне на данни и управляваща информация (команди). В резултат на това данните, които се използват от съвременните информационно-технологични средства са цифрови (или както напоследък често се наричат – дигитални). За разлика от цифровите данни, аналоговите данни не са ограничени до определени числа и представянето им в съответната компютърна система, а зависят от прецизността на уредите за тяхното измерване. Самите данни могат да представляват количествени характеристики от най-различни области на човешката дейност – наука, финанси, медицина, транспорт, физика, химия, природа, време, човешки дейности и др. Интересен случай са данните за компютърните системи от данни или данните за данни.

Големи данни

Големи данни (Big data) е област на информационните технологии, която разглежда методи за анализ и работа с множества от данни, които са твърде сложни и/или твърде големи, за да се извършва от обичайните и масово разпространени компютърни системи. Като предизвикателства за обработката се явяват набирането (придобиване, получаване), съхранението, търсенето, споделянето, анализа и визуализацията на данните. Тенденцията към увеличаване на обема на съвкупностите от данни произтича и от допълнителната информация, която може да се извлече от анализа на простите съвкупности от данни. Големите данни могат да се разглеждат като технология за бизнес-интелигентност с цел съхранение, обработка и анализ на данни, която преди това не е била използвана поради ограниченията на традиционните информационни технологии. Някои от тези множества данни могат да са източник на непознати статистически и функционални зависимости, а други могат да доведат до откриване на непознати до момента грешки или новости/открития (вече има такива в медицината [5]). Предизвикателствата пред големите данни включват регистриране на данни, съхранение на данни, анализ на данни, търсене в данните, споделяне на данни, трансфер на данни, визуализация на данните, запитване в данните, обновяване на данните, сигурност на източниците на данни [6].

Основна особеност на методите в рамките на концепцията на големите данни е възможността за обработка на целия информационен масив за получаване на по-достоверни резултати от анализа. Традиционните методи за анализ на данните (в т.ч. и методите на Извличане на знания от данни - data mining) се основават на ограничена по обем представителна извадка, което закономерно води до увеличаване на грешката. Освен това този подход в много случаи на практика води до увеличени разходи за подготовка на данните, породени от особеностите на представянето им и изискването за тяхното привеждане в определен формат.

Днес, когато се говори за големи данни се има предвид както структурирани така и неструктурирани данни. През последното десетилетие много компании и държавни институции изградиха хранилища за данни и бизнес интелигентни системи с цел да съхраняват и използват тези нарастващи обеми от големи данни. Много световни производители генерират гигантски потоци от данни, които трябва да могат да се анализират за нуждите на бизнеса и пазарите. Няма корпорация или голяма държавна институция, която да не може да се възползва от адекватни решения базирани на обработката и анализа на големите данни. Напротив, липсата на такава възможност би довела до изоставане и регрес спрямо конкурентите или другите институции и законовите рамки в държавите. Ако се разполага с адекватна аналитична платформа за големи данни компаниите могат например да постигнат значително повишаване на ефективността на вземане на решения.

В индустрията се е наложила концепцията за четирите V (или V⁴) като основни показатели на големите данни - Фигура 3.



Фигура 3. Основни показатели на големите данни (V⁴).

Обем (Volume): Големите данни са свързани с голям обем от информация. Получава се натрупване на данни от структурирани и неструктурирани източници (като социалните мрежи и връзките помежду им), мобилните комуникации по 4G/5G мрежите (смарт-телефони, таблети, навигационни системи и AI подобни асистенти), информацията публикувана и споделяна на различни сайтове в Интернет, нарастващия обем на здравни, икономически, производствени, биометрични и множество други данни. Големите Интернет сайтове например следят непрекъснато трафика, потребителите, свалените и качени данни, мултимедийни файлове и други. Това е колосален масив от неструктурирани данни особено при ИТ гиганти като Google, Amazon, Facebook и други от техния ранг. Тези данни могат да се обработват с аналитични средства с цел подпомагане на бизнеса (и на практика това се прави от известно време). Бизнес звената са "потопени" в този нарастващ информационен поток, натрупващ с лекота терабайтове и дори петабайтове от информация.

Скорост (Velocity): Под скорост се разбира темпът на генериране и приемане на данните, а също така и колко бързо дадена организация може да ги анализира и използва. Обработката на данните посредством аналитични модули предоставя възможност за получаване на анализи и отговори във формата на отчети, диаграми, контролни табла и др. Те позволяват да се вземат по-правилни и по-навременни и бързи решения, които повишават ефективността на организацията както и нейните взаимодействия с потребителите и партньорите. Нерядко при обработката на данните са зададени конкретни критични срокове за получаване на резултата. Времевият хоризонт на обработката на данните в реално време може да се окаже твърде кратък – от порядъка на минути, а за процеси, особено чувствителни към времето за реакция като установяване на агресивни опити за измама или биомедицински показатели, където порядъкът може да са секунди. Много пъти за различните приложения на големите данни трябва да се прилага поточен режим, със зададено предварително време за реакция, за да се максимизира полезният резултат.

Разнообразие (Variety): Различните източници на данни като сензори, аудио и видео потоци, транспортни средства, електронна поща, CRM, ERP и други бизнес системи, социални мрежи, продукти свързани със спорта и здравето (wearables) генерират разнородни и разнообразни данни. Големите данни обхващат разновидни типове данни – структурирани и неструктурирани, различно мултимедийно съдържание в различни формати (нови източници – смартфони,

таблети, сензори). Управлението на такъв огромен и сложен набор от данни е комплексна и трудна задача за организациите. Един от подходите за по-добро управление е разделянето на данните чрез аналитични средства - например на основата на произхода или пък на типа и формата на данните. Това позволява по-добро управление на ниво бази данни (СУБД и др.). Много е важно също така да се разграничат новите от старите данни постъпващи от множеството източници.

Достоверност (Veracity): Достоверността касае непълнотата свързана с данните, тяхната наличност, чистота и т.н. Точността и чистотата (липсата на "чужди" несвързани данни) на данните са много сериозен проблем особено при големи обеми данни. Трябва да се обръща специално внимание и за смесването на свързани и несвързани данни, както и за степента на значимост на данните. За да се вземат правилните решения данните трябва да са чисти, консолидирани и консистентни.

Структурирани, полуструктурирани и неструктурирани големи данни

Големите данни, според начина на представяне могат да се разделят на три вида - структурирани, полу-структурирани и неструктурирани – Таблица 1:

Таблица 1. Типове големи данни.

Видове данни	Описание	Примери
Структурирани	Структурираните данни са с ясно дефинирани типове на данните и предоставят възможност за лесно търсене по шаблон (модел на данните)	RFID, NFC, умни измервателни уреди, GPS, сървъри, компютърни приложения, апарати за продажба на стоки, финансови приложения, медицински данни, данни от релационни БД

Таблица 2 - продължение

<i>Видове данни</i>	<i>Описание</i>	<i>Примери</i>
Полу-структурирани	Тези данни обикновено не са с ясно изразената структура на таблични данни с техните дефинирани типове (както е в случая с релационните БД или други таблични данни), но въпреки това съдържат маркери или етикети, които спомагат за групирането и йерархизацията на записите и полетата с данни в тях. Познати са още като самоописващи се структури - структури съдържащи структури. Търсенето по шаблон (модел на данните) при тях е частично затруднено.	Данни от web сайтове, XML или CSV файлове, JSON, AVRO, HTML и др. подобни
Неструктурирани	Неструктурираните данни са без ясно дефинирани типове на данните и обикновено търсенето по шаблон (модел на данните) при тях е доста трудно.	Офис документи, ел. поща, блогове, цифрови изображения, видео, сателитни изображения, аудио/звукови файлове и потоци, някои сензорни данни, сканирани файлове, данни от социални мрежи като Facebook, Twitter и др.

В зависимост от вида на данните начините за тяхната обработка и използване могат да бъдат различни и зависят от наличието на конкретни инструменти и техники за анализ.

Метаданните също са част от големите данни. По същество това са данни за данните и могат да се добиват от анализа на обединени множества от на пръв поглед несвързани източници. Тези данни могат да бъдат превърнати в сериозен източник на информация под формата на факти.

Референтни архитектури за Интернет на обектите

Понятието Референтна Архитектура (РА) служи да обозначи и представи предварително дефиниран модел или множество модели, разработени в даден бизнес и технически контекст и представени с конкретни примери. Референтната архитектура създава стандартна рамка и използва обща терминология, за да опише бизнес, потребителските, функционалните и приложните гледни точки. Общият речник на РА позволява обсъждането на спецификациите за дадено приложение и възможностите за сравнение на референтни архитектури. С помощта на тази терминология се описва една система, дефинициите за нея, блоковете в нея и спецификациите за нейното прилагане. Референтната архитектура предоставя насоки за развитието на една система, решения и архитектури за приложения. Както е дадено в материалите на Международната стандартизираща организация ISO [7], „архитектурата“ представлява "фундаментални концепции или свойства на една система в обкръжаващата я среда, включваща своите елементи, отношения и принципи на дизайн и еволюция". Същият източник определя архитектурна рамка (architecture framework) като "конвенции, принципи и практики за описанието на архитектури, създадени за определена област (домейн) на приложения и/или общество на заинтересовани страни". Тук се обръща внимание и се говори за РА в областта на Интернет на обектите (ИНО).

Важно е да се отбележи, че нивото на абстракция на референтната архитектура е високо, с което се позволява идентифициране и обхващане на най-важните функционалности и особености за всякакво приложение. Като се избягва конкретиката, референтната архитектура може да позволи на следващи проекти да я следват, без проблемите и пречките на излишни и произволни ограничения.

Видове референтни архитектури

Съществуват различни подходи при създаване и представяне на референтни архитектури в областта на ИНО. Въпреки, че повечето са много различни помежду си (има няколко стотин РА), в зависимост от това кой ги е създал, най-общо може да се посочат три класа РА:

- РА на стандартизиращи организации (ISO, IEEE, ITU);
- корпоративни РА (Amazon, Cisco, Cloudera, Intel, Microsoft и други);
- промишлени референтни архитектури.

Ключови функционалности на цялостните Ино референтни архитектури

Може да се извлекат, отличат и представят следните основни функционалности на Ино референтните архитектури от-край-до-край:

- Свързаност и управление на устройствата (обектите) - тази ключова функционалност има за цел да осигури сигурната връзка, надеждната автентикация и лесното управление на разнородните устройства свързани по различаващи се протоколи;
- Интелигентна обработка и анализ по края - тази ключова функционалност позволява обработката на данните да се извършва, доколкото това е възможно и оправдано [8], на границата на Ино архитектурата, за да се прилагат на място бизнес правилата, както и правилата за машинно обучение, което значително намалява времето на процесите по вземане на решения;
- Централизиран анализ и машинно обучение - тази ключова функционалност има за задача да осигури централизираната обработка на Ино данните, да подпомогне машинното обучение и да позволи извличането на важни бизнес зависимости, както и да предложи ефективни действия за оптимизирането на бизнес процесите;
- Приложна и бизнес интеграция - тази ключова функционалност е пряко свързана с интеграцията на Ино архитектурите с бизнес системите и бизнес приложенията в тях като по този начин се осигурява от една страна свързването на оперативните и бизнес процесите, а от друга страна се оптимизира и намалява сложността на интеграционните процеси;
- Сигурност и съответствие от-край-до-край - тази ключова функционалност осигурява сигурността на цялата Ино архитектура и е базирана на средствата за защита, съответствие, автентикация и авторизация, които Ино архитектурата притежава и предоставя за използване.
-

Заклучение

От представените множество референтни архитектури, освен изброените по-горе ключови функционалности е видно, че няма общо мнение за това как и по какъв начин, с какви слоеве и обекти да бъде изградена една референтна

архитектура за Ино. Казано по друг начин, може също така да се заключи, че не съществува най-добра Ино референтна архитектура. Различни Ино референтни архитектури могат да са по-добри при определени обстоятелства и условия за приложение, но по-малко подходящи при други обстоятелства и условия за приложение. Неслучайно има стотици референтни архитектури за приложения, свързани с Ино и няма една наложена се и утвърдена такава.

3. ГЛАВА II. СРЕДСТВА ЗА ОБРАБОТКА НА ГОЛЕМИ ДАННИ И ОПИСАНИЕ НА КОМПОНЕНТИТЕ

Базисни компоненти

Източници на Ино данни

Източниците на данни са компоненти, без които не може (т.е. *sine qua non*) никоя архитектура за Ино. Без самите данни и средствата за тяхното генериране е почти невъзможно да се изграждат и разглеждат следващите нива на архитектурата. Източниците на данни в съвременния свят са от изключително голямо разнообразие както като тип данни (измерване на физични параметри, финансови и борсови индекси, производствени параметри, медицинска информация, умни устройства, транспорт и други), така и като конкретна реализация на сензорите, комуникационните канали и компютърните интерфейси към системите за генериране на данните. Една възможна класификация на приложните области, в които вече се използват източници на Ино данни е дадена на Таблица 3:

Таблица 3. Класификация на източниците на Ино данни.

Приложна област	Подобласти
Приложения за крайни потребители	• Умни къщи • Здравословно хранене и активен спорт • Грижи за по-възрастните
Приложения за компании и организации	• Медицина и здравеопазване • Транспорт и логистика • V2X комуникации (Vehicle-to-Everything) • Строителство и автоматизация • Финансови инструменти

Таблица 2. - продължение

Приложна област	Подобласти
Индустриални приложения	• Производство • Селско стопанство
Инфраструктурни Приложения	• Управление на енергийните ресурси • Наблюдение на околната среда • Приложения на Ино решения за цели градове
Военни приложения	• Internet of Battlefield Things (IoBT) • Ocean of Things
Продуктова дигитализация	• Продукти с възможност за свързване, управление и предоставяне на информация за състоянието

Постъпване на Ино данни

След като данните бъдат генерирани от източниците на данни, същите трябва да постъпят в останалите компоненти на референтната архитектура. Най-често използваните решения за този компонент на архитектурата са Apache NiFi, Eclipse Kura, Streamsets, Azure Data Factory и други.

Разпределяне на събитията за Ино данни

Една от основните характеристики на всички Ино данни е обема и честотата на тяхното генериране и нуждата от обработването на тези големи обеми данни. Най-често използваните решения за този компонент на архитектурата са Apache Kafka, Amazon Kinesis, RABBITMQ, ActiveMQ и други.

Съхранение на Ино данни

Много често се налага получените Ино данни да се съхраняват за пакетни обработки или бъдещо използване. Най-често използваните решения за този компонент на архитектурата са Apache Hadoop, Apache HBase, Apache Kudu, Google BigQuery и други.

Обработка на ИнО данни

След като данните бъдат генерирани, извлечени, почистени и заредени (и съхранени) е време да бъдат извършени аналитичните обработки на тези данни. Най-често използваните решения за този компонент на архитектурата са Apache Hive, Apache Impala, Apache Spark, Amazon Redshift, Azure Synapse Analytics и други.

Визуализация и вземане на решения

Целта на всеки анализ е да предостави информацията необходима за вземане на конкретни решения. Най-често използваните решения за този компонент на архитектурата са Microsoft Power BI, Tableau, Qlik Sense, Google Data Studio и други.

Спомагателни компоненти

Умни крайни устройства

Някои големи корпорации и държавни институции работят с колосален брой ИнО устройства. Използването на умни крайни устройства позволява извършване на агрегиране и трансформация на ИнО данните и частичната им обработка на място, което допринася за по-малкото натоварване на комуникациите и по-бързата последваща обработка.

Облачна телеметрия и трансформация

Телеметрията е процес на автоматична комуникация с множество източници на данни. Данните постъпващи от телеметрията се използват за наблюдение на различни параметри на устройствата и за реструктуриране, комбиниране и обработка на постъпващите от ИнО устройствата данни.

Машинно обучение

Машинното обучение е изучаването на компютърните алгоритми с цел автоматичното им оптимизиране посредством натрупания опит. Машинното обучение позволява създаването и изпълнението на предсказващи алгоритми на базата на натрупаните исторически данни.

Управление и контрол на достъпа

Възможността за управление и контрол на потребителите позволява функционално и йерархично разпределяне на достъпа, отговорностите и ролите на различните потребители на Ино РА. Управлението на потребителите позволява и създаване на класове от потребители с различни нива на достъп и възможности за взаимодействие с Ино РА.

Проследяване и записване

Воденето на журнали (logs) е също е критичен елемент от архитектурата. По този начин се обезпечават средства за наблюдение и отстраняване възникващи проблеми и откази на отделни подсистеми, на цели приложения както и на цялата архитектура.

Висока надеждност и наличност (резервираност)

Всички критични Ино РА трябва да разполагат със средства за възстановяване на услугите и приложенията след загуба на услугите или хардуерни проблеми. Принципите, на които се стъпва, обикновено са заимствани от стандартните ИТ решения за възстановяване след бедствени събития (Disaster Recovery) и системи за бързо възстановяване.

Общи изисквания

Сигурност

Сигурността е неотменен компонент в днешните ИТ системи. РА не правят изключение. При РА този компонент е част от цялата система. Трябва да се подчертае, че сигурността на РА е много обширна и комплексна тема, която сама по себе си би могла да е обект на задълбочени изследвания и разработки, които са извън обхвата на настоящия труд.

Управление

Управлението е друг неотменен компонент в днешните ИТ системи и като такъв е задължителен елемент на Ино РА. Управлението на РА е също много

обширна и комплексна тема, която сама по себе си би могла да е обект на задълбочени изследвания и разработки, които са извън обхвата на настоящия труд.

Изисквания към компонентите на референтни архитектури за Ино

При подбора на компонентите за реализация на Ино РА са взети предвид характеристиките изброени по-долу. Трябва да се отбележи, че това не са всички възможни характеристики на елементи за изграждане на Ино РА - съществуват и много други такива, но за целите на настоящия труд, представените тук характеристики на компоненти на Ино РА дават възможност за осъществяване на работеща и завършена архитектура за Ино.

Мрежови комуникации: Ино системите по самата си същност са силно зависими от мрежовите комуникации.

Протоколи за обмен на данни: Комуникационните протоколи, които Ино РА използват също са доста разнородни в зависимост от различните типове мрежи.

Силно разпределени системи: Ино могат да обхващат цели сгради, градове или по-големи области.

Надеждност: Под надеждност в Ино архитектура се разбира при работата на архитектурата да изпълняват стабилно заданията, които са ѝ поставени.

Управляемост: Много Ино устройства, мрежи и системи работят автономно и независимо. Въпреки това, трябва да е възможно да се осигури управлението на всички компоненти на Ино архитектурата дори и в случай на отказ на отделни системи, подсистеми или елементи.

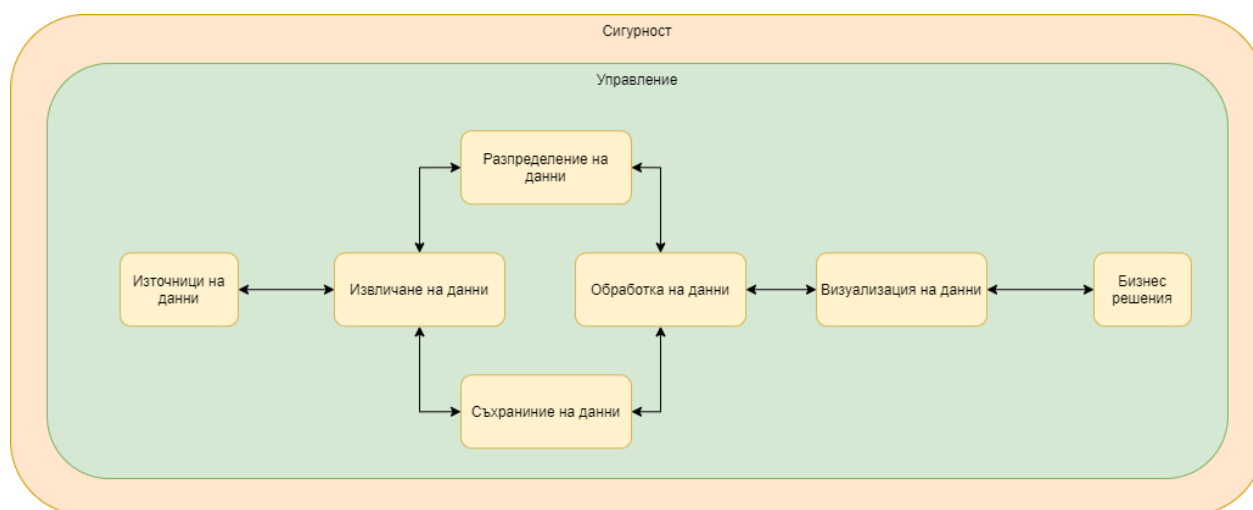
Модулност: Компонентите на Ино архитектура трябва да притежават характеристиката модулност, за да могат да бъдат комбинирани заедно в различни конфигурации според нуждите и задачите поставени пред Ино архитектурата.

Работа в реално време: Често се налага Ино система да работи в реално или почти реално време - Ино данните и събитията формират постоянни информационни потоци, които трябва да се предадат, обработят и визуализират.

Мащабиране: Под мащабиране в Ино архитектура се разбира възможността на архитектурата да извършва своите функции и да постига своите задачи дори и при нарастване на сложността и обема на постъпващите Ино данни.

Масштабирането може да има и обратна посока, с цел икономии на ресурси и средства, ако изискванията към РА намалеят.

На базата на гореизложеното, е представена принципна схема на Ино РА (Фигура 4), която включва базисните, спомагателните и общите компоненти на архитектура за Ино приложения.



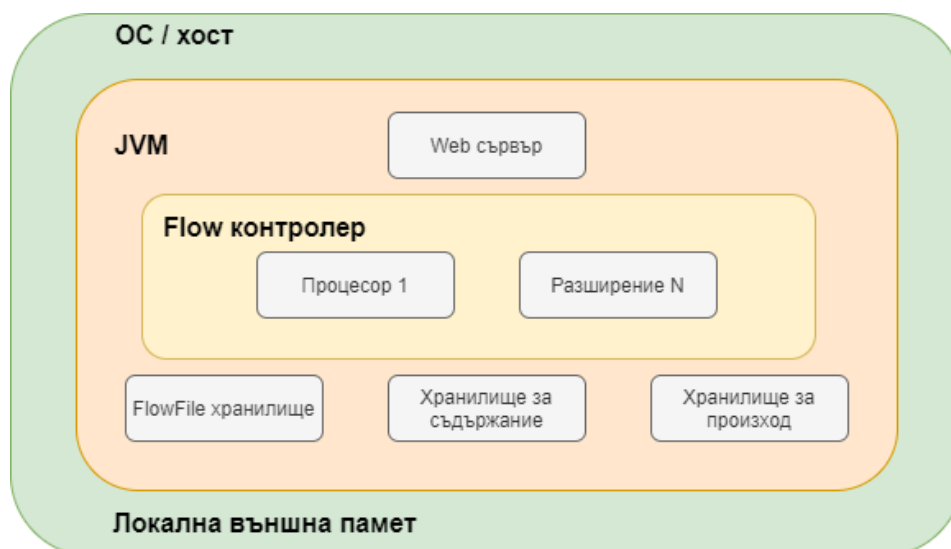
Фигура 4. Принципна схема на Ино РА.

Избрани ключови компоненти за изграждане на референтна архитектура за Ино

Apache NiFi

Apache NiFi [9] е избран за компонент за приемане на данни в предложената Ино архитектура. Apache NiFi предоставя гъвкави възможности за приемане на разнородни данни в сравнение с алтернативните възможности (Eclipse Kura, Streamsets, Azure Data Factory), интегрира се сравнително лесно с другите компоненти от Apache Ино екосистемата като Apache Hive, Apache Impala, Apache Kafka и други., и притежава функционалности за първична обработка на постъпващите Ино данни.

На практика NiFi е Java приложение, което се изпълнява във виртуална Java машина (JVM), която от своя страна работи върху операционната система на сървъра. Основните компоненти на NiFi са представени на Фигура 5.

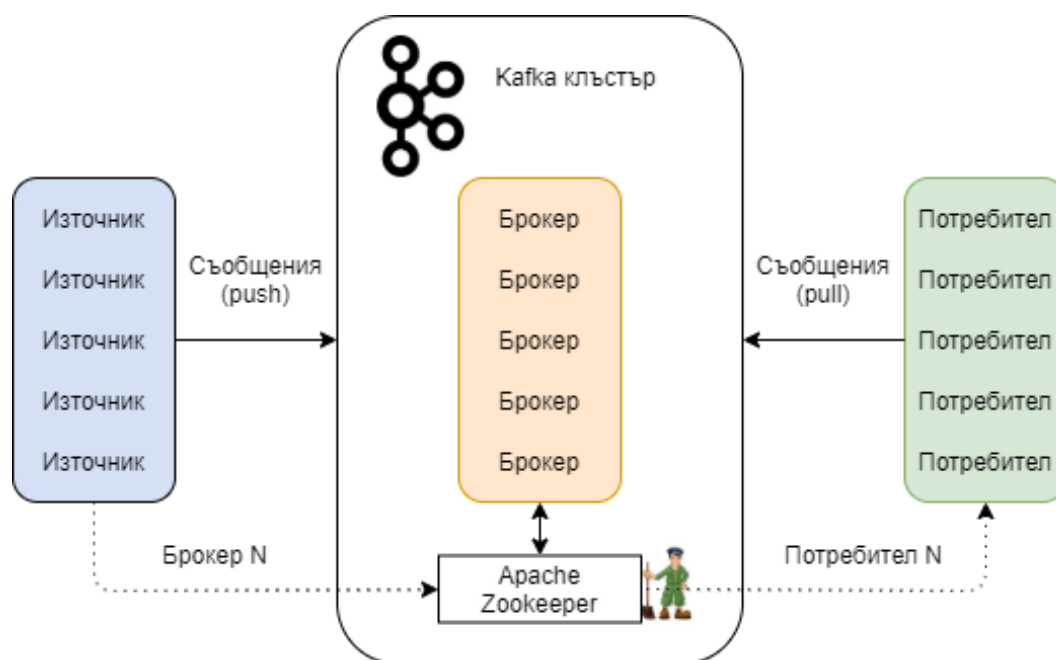


Фигура 5. Архитектура на Apache NiFi.

Apache Kafka

Apache Kafka [10] е избран за компонент за разпределяне на събитията в предложената ИнО архитектура. В сравнение алтернативите (Amazon Kinesis, RA1TMQ, ActiveMQ и други) Apache Kafka осигурява висока съвместимост с другите компоненти от Apache ИнО екосистемата, има големи възможности за мащабиране и много силна софтуерна поддръжка и развитие.

В съчетание с вградените в Apache Kafka механизми за предотвратяване на загуба на данни (zero data loss), наличност на услугите при възстановяване от срив и, не на последно място, средства за защита и сигурност на предаване на данните, възможностите за лесна интеграция с Spark Streaming, HBase, Cloudera Manager и др. Apache Kafka е не случайно избрана среда за основа на ИнО архитектурите позволяваща лесна и скалируема обработка на големи обеми данни - Фигура 6.

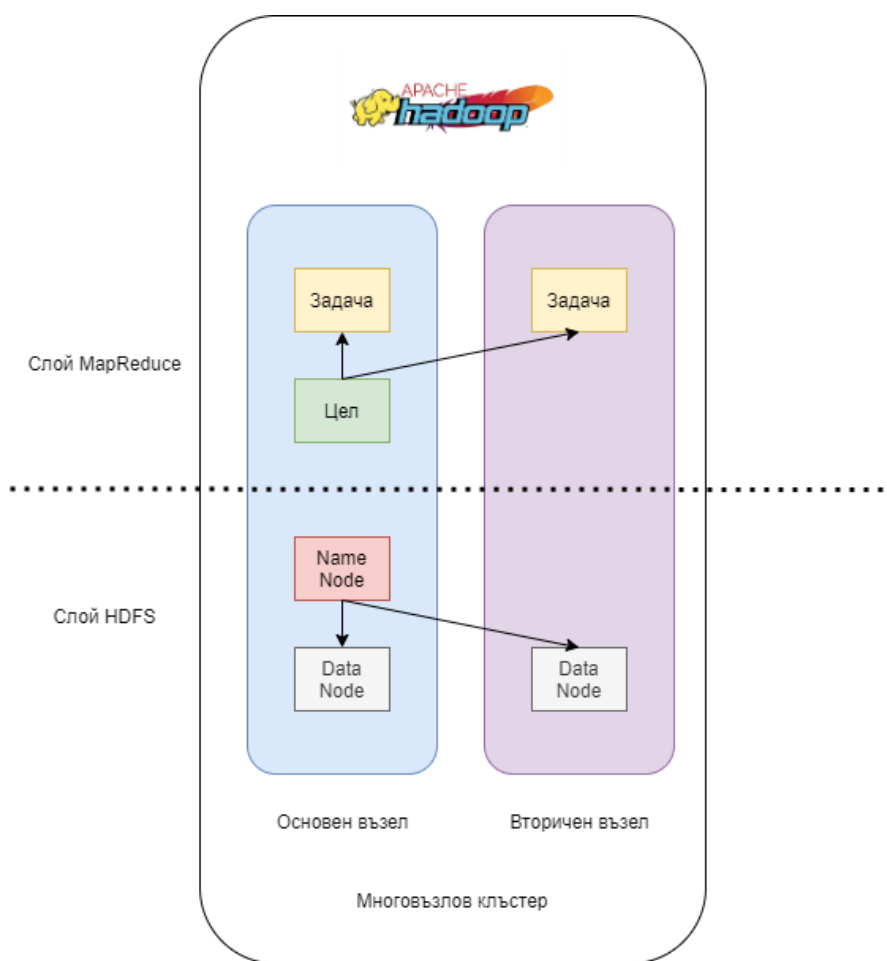


Фигура 6. Архитектура на Apache Kafka.

Apache Hadoop

Apache Hadoop [11] е избран за компонент за съхранение на данните в предложената ИНО архитектура. В сравнение алтернативите (Apache HBase, Apache Kudu, Google BigQuery и други) Apache Hadoop, подобно на Apache Kafka, осигурява висока съвместимост с другите компоненти от Apache ИНО екосистемата, има големи възможности за мащабиране и много силна софтуерна поддръжка и развитие.

Apache Hadoop се състои най-общо от два функционални слоя (MapReduce и HDFS) и клъстер с много възли - Фигура 7.

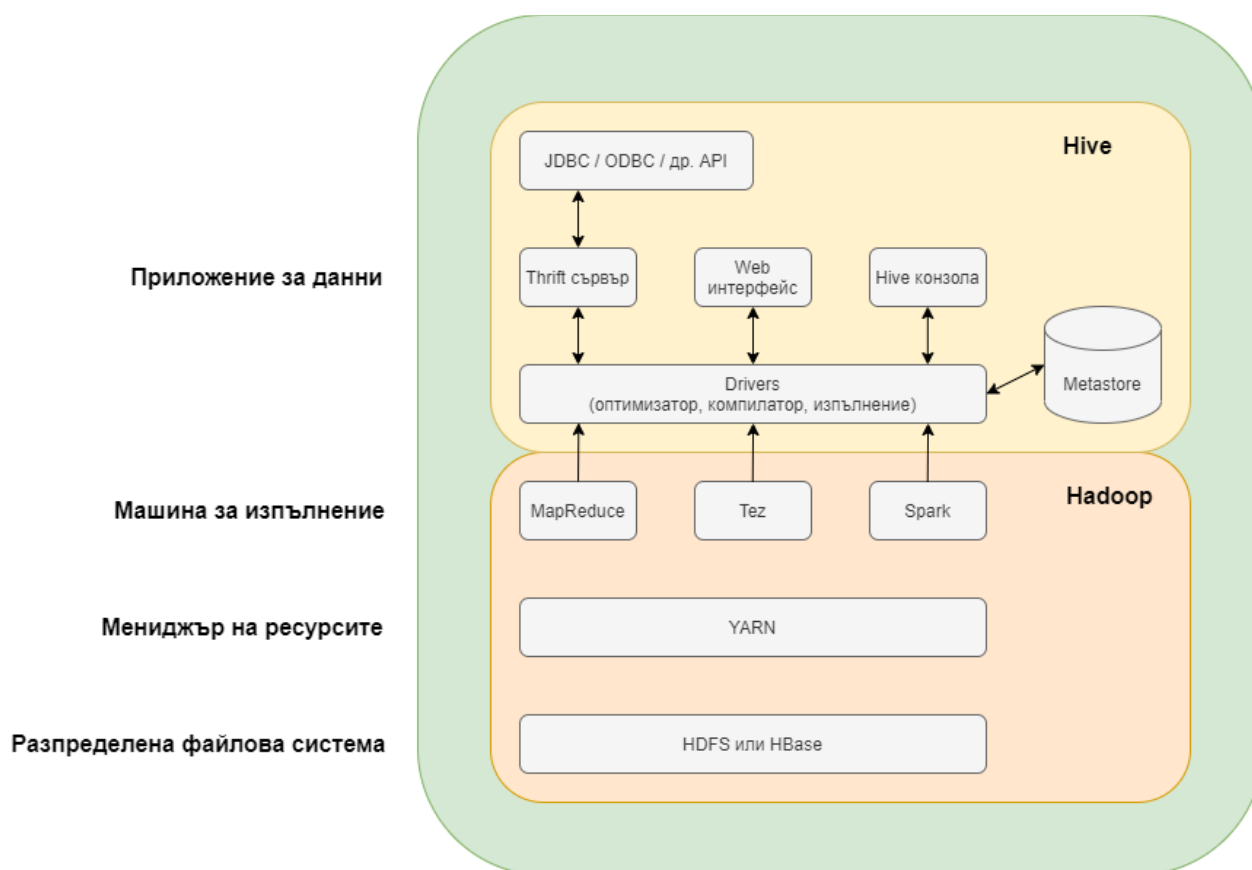


Фигура 7. Архитектура на Apache Hadoop.

Apache Hive

Един от избраните компоненти за обработка и анализ на данните в предложената ИНО архитектура е Apache Hive [12]. Другият избран компонент за обработка и анализ на данните в предложената ИНО архитектура е Apache Impala [13]. Двата обработващо-аналитични продукта са сравнени в приложение А. Както и повечето елементи от екосистемата на Apache, Apache Hive осигурява висока съвместимост с останалите Apache ИНО компоненти, като има големи възможности за мащабиране и много солидна софтуерна поддръжка и развитие.

Apache Hive поддържа естествено анализ на големи набори от данни записани в Hadoop HDFS. Apache Hive е съвместим и други файлови системи като Amazon Simple Storage System (S3) файловата система и Alluxio. Посредством HiveQL могат да се трансформират лесно заявки към MapReduce, Apache Tez и Spark задачи. И трите компонента са съвместими и използват услугите на Hadoop YARN (Yet Another Resource Negotiator) - Фигура 8.

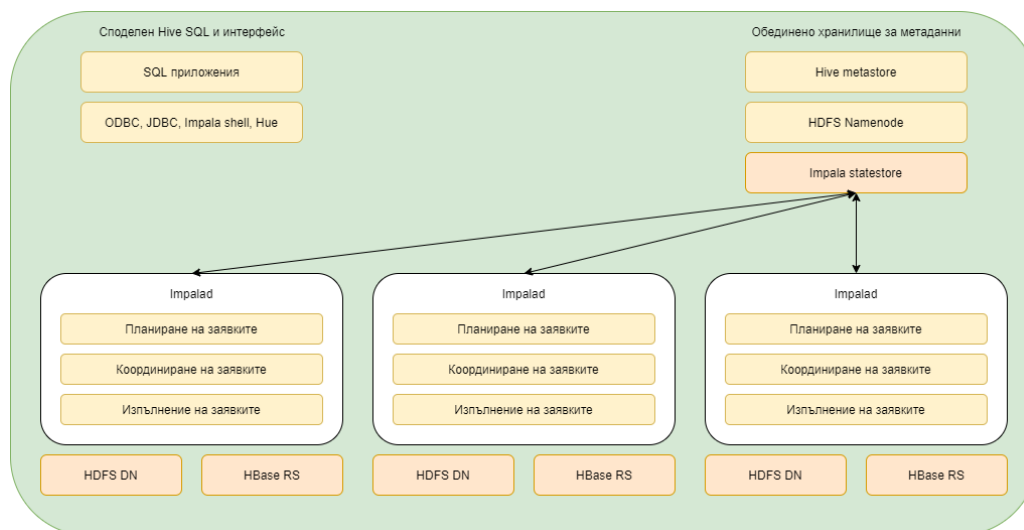


Фигура 8. Архитектура на Apache Hive.

Apache Impala

Вторият избран компонент за обработката на ИНО данните е Apache Impala [13]. Това е софтуерен проект, който осигурява висока производителност при ниска латентност на изпълняваните SQL заявки към данните записани в най-често използваните файлови формати от екосистемата на Apache Hadoop. Бързият отговор на тези SQL заявки предоставя възможности за интерактивно изследване на данните и допълнителна фина настройка на аналитичните заявки. Тук под интерактивност се има предвид възможността за получаване на бързи отговори на заявките т.е. инженерите, аналитиците и учените работещи с големи данни могат в реално или почти реално време да видят резултатите от своите заявки. Това силно контрастира с обичайния подход за пакетна (бач) обработка на SQL заявки към Hadoop хранилища, които са бавни и не могат да се използват интерактивно.

Основните компоненти в архитектурата на Apache Impala са представени на Фигура 9.

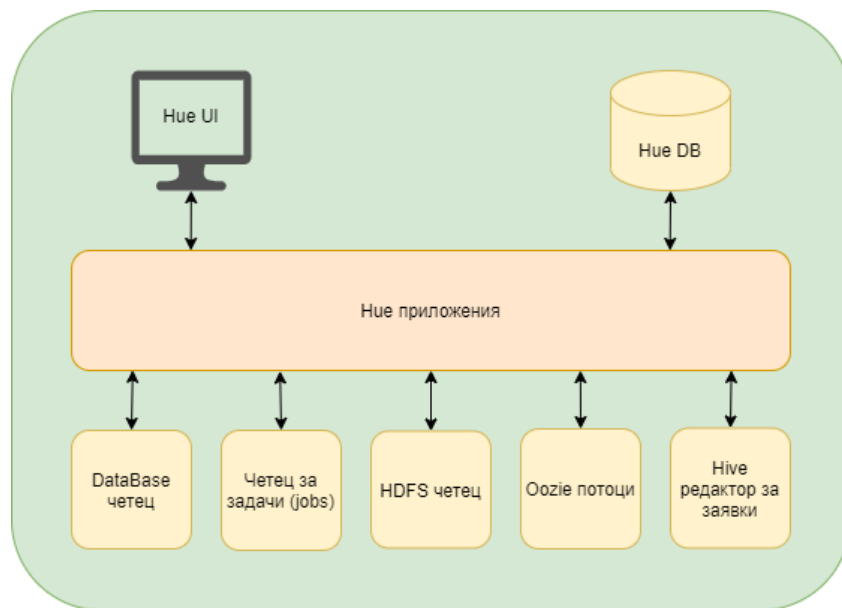


Фигура 9. Архитектура на Apache Impala.

Hue SQL редактор

За интерфейс към аналитичните машини на Apache Hive и Apache Impala е избран Hue [14] SQL редактор. Алтернативните варианти за работа с ИНО данните са конзолните инструменти на Apache Hive (beeline) и Apache Impala (impala-shell). Hue е достъпен за работа директно през браузър, а работата с конзолните инструменти предполага достъп до някой от възлите на клъстера на архитектурата и умения за работа в Linux среда, като това невинаги е възможно или удобно. Hue е с отворен код, разпространява се в съответствие с Apache License 2.0 и е проектиран преди всичко за работа в разпределени и облачни среди. Чрез Hue могат лесно да се изпълняват SQL-подобни заявки към различни бази от данни и хранилища за данни. Целта му е да популяризира и разпространи писането на заявки за тези сценарии. Hue се разпространява посредством версии публикувани на официалния му сайт. Hue е включен като компонент в Cloudera Data Platform (CDP) и в Hadoop услугите предоставяни от облачните доставчици като Amazon AWS, Google Cloud Platform (GCP) и Microsoft Azure.

На Фигура 10 са представени основните компоненти и типичното приложение на Hue SQL редактора в ИНО среда.

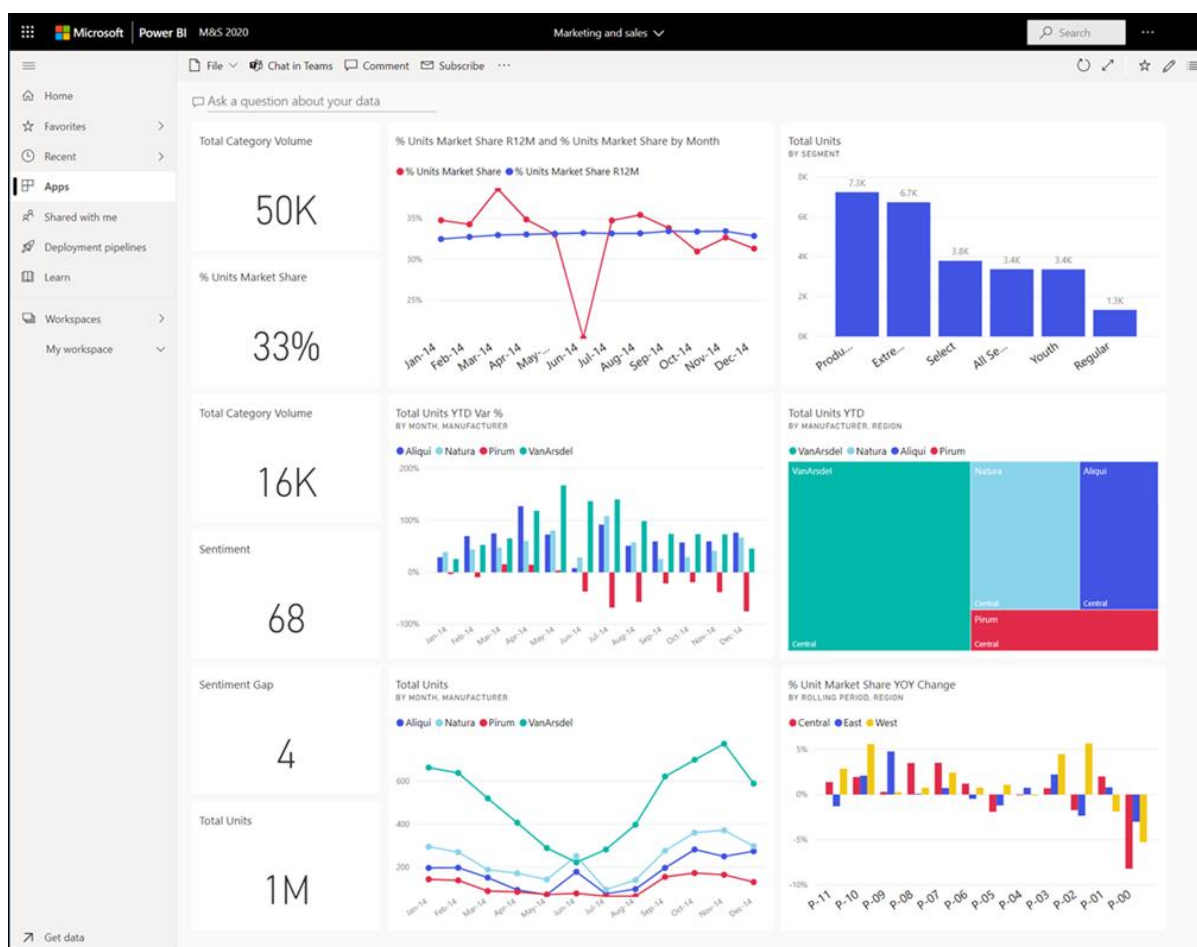


Фигура 10. Hue компоненти и типични приложения.

Microsoft Power BI

Microsoft Power BI [15] е избран за компонент за визуализация в предложената ИНО архитектура. В сравнение с алтернативите (Tableau, Qlik Sense, Google Data Studio и други) Microsoft Power BI лесно се инсталира и интегрира с останалите компоненти от архитектурата. Microsoft Power BI е мощен инструмент, достъпен за академичните общности и разчита на софтуерния гигант Microsoft за поддръжка и развитие.

Microsoft Power BI предоставя облачно базирани услуги за бизнес интелигентност (business intelligence - BI) познати като Power BI Services заедно с десктоп-базиран потребителски интерфейс познат като Power BI Desktop. Microsoft Power BI Embedded е част от Microsoft Azure облачната платформа. Чрез Microsoft Power BI могат да се използват хранилища от данни като функционално са включени: подготовка на данните, откриване на данните и интерактивни информационни табла. Microsoft Power BI позволява създаването и зареждането на визуализации създадени от потребителите - Фигура 11.



Фигура 11. Визуализация с Microsoft Power BI. [15]

Заклучение

Съществува голямо разнообразие от средства за обработка на големи данни са. В тази глава бяха разгледани базисните, спомагателните и общите компоненти използвани за обработка на данни както и изискванията, на които те трябва да отговарят. Също така, бяха описани ключовите компоненти за изграждане на една референтна архитектура за ИНО. Накрая бяха представени избраните компоненти, които са използвани при проектирането на модела за референтна архитектура за ИНО описан в следващата глава.

4. Глава III. МЕТОД ЗА ОЦЕНКА НА РЕФЕРЕНТНИ АРХИТЕКТУРИ. РАЗРАБОТВАНЕ НА МОДЕЛИ ЗА ИНО РА.

Дефиниране на метод за анализ и оценка на референтни архитектури

Както бе посочено в I глава, през последните години много организации и компании създадоха свои собствени ИНО референтни архитектури. Всяка РА е

фокусирана върху конкретни пазарни сегменти, производители и доставчици и/или схващанията на авторите на РА. Затова е трудно за екип натоварен да разработи нова Ино РА за конкретен бизнес сценарий да оцени всяко едно възможно решение от вече създадените и налични такива. Ето защо би било целесъобразно и полезно да се разработи метод за количествена оценка на съществуващи вече Ино референтни архитектури, който да даде максималните стойности за очакваната приложимост, ефективност и съответствие на референтната архитектура. Разработен е именно такъв метод, който е наречен "Еталонна РА". Използването на този метод позволява преодоляването на описаната трудност при създаване на нови архитектури, защото те вече могат да се оценяват и позиционират спрямо заложената в метода стойност за "Еталонна РА".

Методът се основа на експертна оценка за тежестта и стойността на критериите от частта "Главни характеристики на Ино системи" на стандарта ISO 30141. Стандартът посочва девет нива (групи) като всяко от тези нива може да има по едно или повече поднива (критерии). Всяко подниво се оценява с теглови коефициент от 1 до 5 в зависимост от експертната оценка за важността на това ниво в глобален аспект. Степента на съответствие на дадена конкретна Ино РА спрямо разглеждания критерий (подниво) се оценява със стойност от 0 до 9. Оценката за всяка група (ниво) е комплексна и се получава чрез сумиране на отделните тегловите коефициенти (1-5) и стойностите (0-9) за всяко подниво в тази група.

ISO стандартът "ISO/IEC 30141:2018(E), Information technology – Internet of Things Reference Architecture" [16] е избран поради факта, че от всички проучени и изследвани източници за анализ и оценка на РА този стандарт включва най-пълнен и изчерпателен набор от критерии.

В Таблица 4 са представени критериите за оценка на референтни архитектури за Ино заложи в ISO/IEC 30141:2018(E) и е направена оценка на две конкретни компании (X и Y). Тегловите коефициенти за оценка, както и конкретните стойности на критериите за анонимните компании X и Y са получени от допитване до експерти в областта на Ино РА.

Таблица 4. Количествена оценка на референтни архитектури за Ино.

Критерии за оценка на Ино Ре- ферентни Архитектури	Тегло на кое- фициента в ИНО РА (1-5)	Компания Х		Компания У	
		Стойност на критерия	Комплексна стойност	Стойност на критерия	Комплексна стойност
1 Системни Ино характеристики			163		203
1.1 Автоматична конфигурация	5	8	40	7	35
1.2 Разделение между функцио- нални и управляващи ресурси	3	5	15	9	27
1.3 Високо ниво на разпределение	2	5	10	7	14
1.4 Мрежови комуникации	5	5	25	9	45
1.5 Управление и работа на мре- жата	2	5	10	8	16
1.6 Възможности за работа в ре- ално време	3	6	18	7	21
1.7 Автоматично описание	4	0	0	0	0
1.8 Абонамент за услугите	5	9	45	9	45
2 Характеристики на Ино услугите			16		18
2.1 Информация за съдържанието	2	0	0	0	0
2.2 Информация за контекста	2	0	0	0	0
2.3 Темпоралност	2	8	16	9	18
3 Характеристики на Ино компо- нентите			114		150
3.1 Заменяемост	4	7	21	7	28
3.2 Откриваемост	4	0	0	0	0
3.3 Модулност	3	7	21	7	21
3.4 Мрежова свързаност	5	5	25	9	45
3.5 Споделяне	4	5	20	7	35
3.6 Уникална идентификация	4	5	20	7	28
4 Съвместимост			63		63
4.1 Поддръжка на предходни техно- логии	4	7	28	7	28
4.2 Добре дефинирани компоненти	5	7	35	7	35
5 Използваемост			81		81
5.1 Гъвкавост	4	9	36	9	36
5.2 Управляемост	5	9	45	9	45

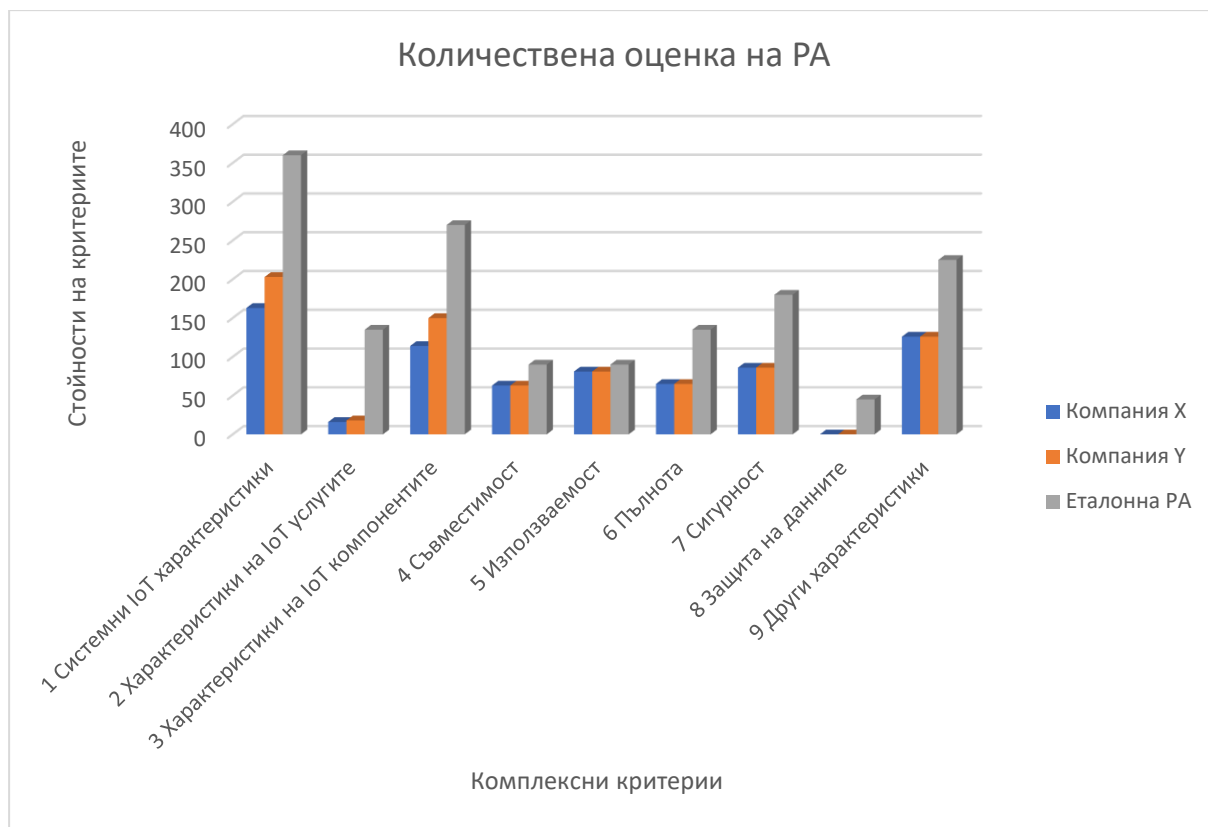
Таблица 3 - продължение

Критерии за оценка на Ино Ре- ферентни Архитектури	Тегло на кое- фициента в ИНО РА (1-5)	Компания Х		Компания Y	
		Стойност на критерия	Комплексна стойност	Стойност на критерия	Комплексна стойност
6 Пълнота			65		65
6.1 Достоверност	5	5	25	5	25
6.2 Надеждност	4	5	20	5	20
6.3 Устойчивост	4	5	20	5	20
7 Сигурност			86		86
7.1 Наличност	4	9	36	9	36
7.2 Конфиденциалност	5	5	25	5	25
7.3 Интегритет	4	5	20	5	20
7.4 Безопасност	1	5	5	5	5
8 Защита на личните данни	3	0	0	0	0
9 Други характеристики			126		126
9.1 Данни - обем, скорост, досто- верност, променливост и разнооб- разие	5	5	25	5	25
9.2 Хетерогенност	4	5	20	5	20
9.3 Съответствие с нормативните уредби	5	5	25	5	25
9.4 Машабируемост	4	9	36	9	36
9.5 Доверителност	4	5	20	5	20

Прилагайки този метод референтната Ино архитектура на компанията Х се оценява на 714 точки, а референтната Ино архитектура на компанията Y се оценява на 792 точки. Максимално достижимата стойност на Еталонната референтна архитектура се равнява на 1530 точки. Графичното представяне на сравнителната оценка на Ино референтни архитектури е дадено на Фигура 12.

Графичната оценка е фокусирана върху общата оценка на критериите и съдържа за сравнение стойностите за Еталонната РА, като оценки на най-добрата възможна Ино РА.

Разработеният метод за количествена оценка на референтни архитектури за Ино е публикуван в [17].



Фигура 12. Графично представяне на оценките на РА.

Използване на метода за анализ и оценка на референтните архитектури на Microsoft, Cloudera и Google

След като бе проверена принципната приложимост на метода за оценка на референтни Ино РА архитектури за компаниите X и Y естественото продължение на изследването е да се направи и оценка на неанонимни Ино РА. За целта са избрани разработените Ино РА на едни от най-големите и известни компании на пазара на Ино РА, които заемат ключов дял от този пазар като доставчици на такива услуги към момента [18]. Това са Ино РА на Microsoft, Cloudera и Google. В Таблица 5 е представена количествена оценка на съответната референтна архитектура.

Таблица 5. Оценка на конкретни референтни архитектури за Ино.

Критерии за оценка на Ино Референтни Архитектури	Microsoft Azure		Cloudera CDH6		Google GCP	
1 Системни Ино характеристики		Оценка		Оценка		Оценка
1.1 Автоматична конфигурация	Посредством Ино Hub модул	8	Посредством Cloudera Manager	8	Spring Cloud GCP	7
1.2 Разделение между функционални и управляващи ресурси	Посредством Identity/Twin модул	5	Посредством Cloudera Manager	9	GCP Key Management Service (KMS)	8
1.3 Високо ниво на разпределение	Посредством HDInsight модул	5	Посредством CHD 6	7	Вградено в платформата	9
1.4 Мрежови комуникации	Посредством Event Hub модул	5	Посредством Cloudera Manager	9	GCP Virtual Private Cloud (VPC)	8
1.5 Управление и работа на мрежата	Посредством Event Hub модул	5	Посредством Cloudera Manager	8	GCP Virtual Private Cloud (VPC)	8
1.6 Възможности за работа в реално време	Посредством HDInsight модул	6	Cloudera Stream Processing (CSP)	7	Fluentd / BigQuery	7
1.7 Автоматично описание	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0
1.8 Абонамент за услугите	Част от Azure платформата	9	Част от Cloudera платформата	9	Част от GCP платформата	9
2 Характеристики на Ино услугите						
2.1 Информация за съдържанието	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Част от GCP платформата	9
2.2 Информация за контекста	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Част от GCP платформата	9
2.3 Темпоралност	Посредством NTP модул	8	Посредством Cloudera Manager	9	Част от GCP платформата	9

Таблица 4 - продължение

Критерии за оценка на Ино Реперентни Архитектури	Microsoft Azure		Cloudera CDH6		Google GCP	
3 Характеристики на Ино компонентите		Оценка		Оценка		Оценка
3.1 Заменяемост	В зависимост от производителя	7	В зависимост от производителя	7	В зависимост от производителя	7
3.2 Откриваемост	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	2	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0
3.3 Модулност	В зависимост от производителя	7	В зависимост от производителя	7	В зависимост от производителя	7
3.4 Мрежова свързаност	Посредством Event Hub модул и стандартни мрежови протоколи	5	Посредством Cloudera Manager	9	GCP Virtual Private Cloud (VPC)	8
3.5 Споделяне	Ино Central Application Export модул	5	Cloudera Shared Experience SDX	7	API	5
3.6 Уникална идентификация	В зависимост от производителя	7	В зависимост от производителя	7	В зависимост от производителя	5
4 Съвместимост						
4.1 Поддръжка на предходни технологии	В зависимост от производителя	7	В зависимост от производителя	7	В зависимост от производителя	7
4.2 Добре дефинирани компоненти	В зависимост от производителя	7	В зависимост от производителя	7	В зависимост от производителя	7
5 Използваемост						
5.1 Гъвкавост	Вградено в платформата	9	Вградено в платформата	9	Вградено в платформата	9
5.2 Управляемост	Вградено в платформата	9	Cloudera Manager	9	Вградено в платформата	9

Таблица 4 - продължение

Критерии за оценка на Ино Ре- ферентни Архитектури	Microsoft Azure		Cloudera CDH6		Google GCP	
6 Пълнота		Оценка		Оценка		Оценка
6.1 Достоверност	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
6.2 Надеждност	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
6.3 Устойчивост	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
7 Сигурност						
7.1 Наличност	Вградено в плат-формата	9	Вградено в платформата	9	Вградено в плат-формата	9
7.2 Конфиденциалност	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
7.3 Интегритет	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
7.4 Безопасност	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
8 Защита на личните данни	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Не е налично по подразбиране. Възможно е създаване на програмни модули за целта.	0	Вградено в платформата	9

Таблица 4 - продължение

Критерии за оценка на Ино Реперентни Архитектури	Microsoft Azure		Cloudera CDH6		Google GCP	
9 Други характеристики		Оценка		Оценка		Оценка
9.1 Данни - обем, скорост, достоверност, променливост и разнообразие	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
9.2 Хетерогенност	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
9.3 Съответствие с нормативните системи	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5
9.4 Масшабируемост	Вградено в платформата	9	Вградено в платформата (Cloudera Data Platform)	9	Вградено в платформата	9
9.5 Доверителност	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5	В зависимост от конкретната реализация на архитектурата	5

Графично представяне на количествените оценки на референтните архитектури на Microsoft, Cloudera и Google е представено на Фигура 13.



Фигура 13. Оценка на Ино РА за Microsoft, Google и Cloudera.

Прилагайки този метод референтната Ино архитектура на Microsoft се оценява на 722 точки, на референтната Ино архитектура на Cloudera се оценява на 805 точки, а на референтната Ино архитектура на Google на 812 точки. Максимално достижимата стойност на Еталонната референтна архитектура се равнява на 1530 точки. Представената графична оценка е фокусирана върху общата оценка на критериите и съдържа за сравнение стойностите за Еталонната РА, като оценки на най-добрата възможна Ино РА.

Моделиране с компоненти за референтна архитектура за Ино

Създаден е модел на компонентна референтна архитектура за Ино, който включва следните пакети:

- Apache NiFi;
- Apache Kafka;

- Apache Hadoop (HDFS);
- Apache Hive;
- Apache Impala;
- Hue, impala-shell, beeline;
- Microsoft Power BI;
- Всички те работят в интеграция с Cloudera CDH6.

На базата на изброените компоненти, тяхната функционалност и характеристики разгледани в Таблица 4 може да се направи обща оценка за предложената архитектура. Тази оценка е представена в Таблица 6, а нейната стойност е 816.

Таблица 6. Количествена оценка на предложената архитектура.

Критерии за оценка на Ино Реперентни Архитектури	Тегло на коефициента в ИНО РА (1-5)	Предложена архитектура	
		Стойност на критерия	Комплексна стойност
1 Системни Ино характеристики			180
1.1 Автоматична конфигурация	5	8	40
1.2 Разделение между функционални и управляващи ресурси	3	8	24
1.3 Високо ниво на разпределение	2	6	12
1.4 Мрежови комуникации	5	5	25
1.5 Управление и работа на мрежата	2	5	10
1.6 Възможности за работа в реално време	3	8	24
1.7 Автоматично описание	4	0	0
1.8 Абонамент за услугите	5	9	45
2 Характеристики на Ино услугите			36
2.1 Информация за съдържанието	2	5	10
2.2 Информация за контекста	2	5	10
2.3 Темпоралност	2	8	16
Критерии за оценка на Ино Реперентни Архитектури	Тегло на коефициента в ИНО РА (1-5)	Предложена архитектура	
		Стойност на критерия	Комплексна стойност
3 Характеристики на Ино компонентите			142
3.1 Заменяемост	4	7	21
3.2 Откриваемост	4	2	8

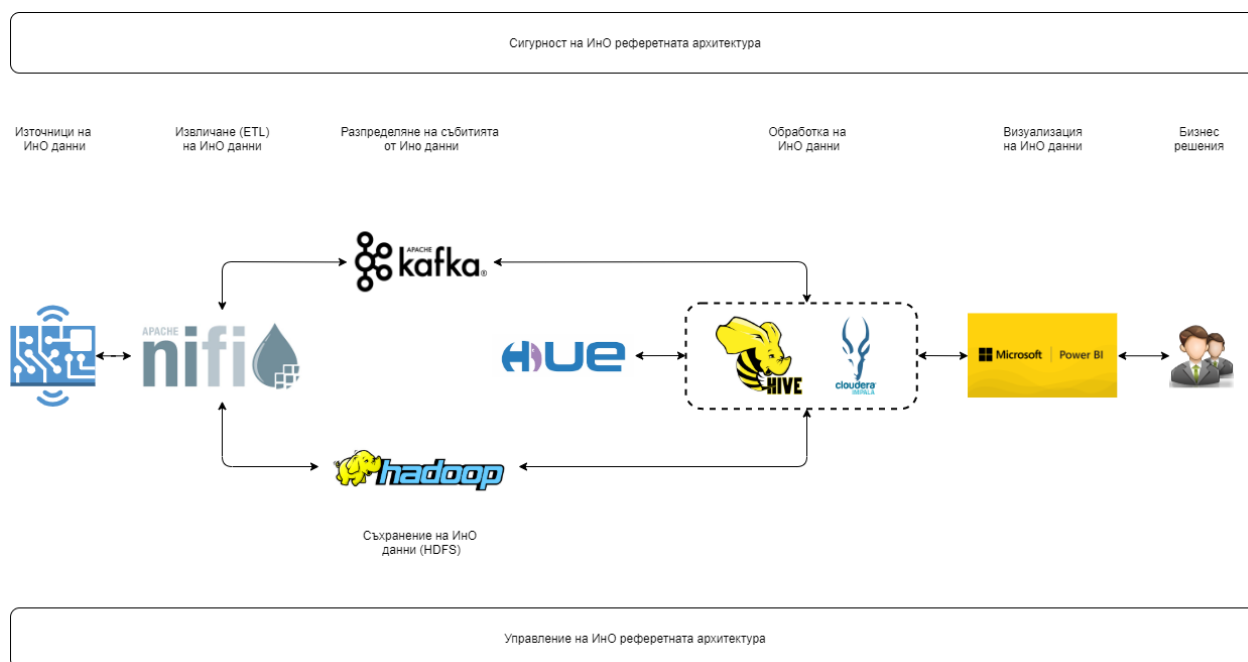
Таблица 5 - продължение

Критерии за оценка на Ино Ре- ферентни Архитектури	Тегло на кое- фициента в ИНО РА (1-5)	Предложена архитектура	
		Стойност на критерия	Комплексна стойност
3.3 Модулност	3	7	21
3.4 Мрежова свързаност	5	8	40
3.5 Споделяне	4	5	20
3.6 Уникална идентификация	4	8	32
4 Съвместимост			63
4.1 Поддръжка на предходни техно- логии	4	7	28
4.2 Добре дефинирани компоненти	5	7	35
5 Използваемост			72
5.1 Гъвкавост	4	8	32
5.2 Управляемост	5	8	40
6 Пълнота			81
6.1 Достоверност	5	5	25
6.2 Надеждност	4	7	28
6.3 Устойчивост	4	7	28
7 Сигурност			86
7.1 Наличност	4	9	36
7.2 Конфиденциалност	5	5	25
7.3 Интегритет	4	5	20
7.4 Безопасност	1	5	5
8 Защита на личните данни	3	0	0
9 Други характеристики			156
9.1 Данни - обем, скорост, досто- верност, променливост и разнооб- разие	5	7	35
9.2 Хетерогенност	4	6	24
9.3 Съответствие с нормативните уредби	5	5	25
9.4 Машабируемост	4	9	36
9.5 Доверителност	4	9	36

Количествената оценка от 816 показва, че предложената архитектура е максимално изчистена и функционално пълна без да поставя сериозни изисквания за

използваните ресурси тъй като е базирана на Open Source компоненти. В същото време тази архитектура не прави компромиси с качеството, гъвкавостта, мащабирането, надеждността и т.н., което е видно от Таблица 5.

Принципна схема на предложената компонентна референтна архитектура за ИНО е представена на Фигура 14.



Фигура 14. Принципна схема на предложената компонентна ИНО РА.

Разработеният модел притежава основните компоненти необходими за изграждането на ИНО РА. Характерна особеност на предложения модел е включването на компонента Apache NiFi. Това значително подобрява възможностите за интеграция с разнородни и разнообразни източници на ИНО данни и другите компоненти на ИНО РА. Добавянето на Apache NiFi в предложения модел за ИНО РА води до съществено предимство на модела спрямо разгледаните и проучените съществуващи ИНО РА.

Заклучение

Разработен е метод за количествена оценка на референтни архитектури за ИНО [17]. Проверена е приложимостта на този метод за конкретни референтни архитектури за ИНО и е извършена количествена оценка на тези референтни архитектури.

Извършен е анализ на възможните начини за създаване на референтни архитектури за Ино и е избран компонентния подход за израждане на референтни архитектури за Ино.

Предложена е гъвкава компонентна референтна архитектура за Ино. Същата е успешно реализирана на два компютърни клъстера:

- учебен тестов 3 възлов клъстер;
- мощен производствен 40 възлов компютърен клъстер.

5. Глава IV. ПРИЛОЖЕНИЕ НА СЪЗДАДЕНАТА РА ЗА ОБРАБОТКА И АНАЛИЗ НА РАЗЛИЧНИ ТИПОВЕ ДАННИ

Тестване на референтната архитектура в ръчен режим на зареждане на данни

Тестване с метеорологични Ино данни

Източник на метеорологични Ино данни

Първите представени примери за тестване на референтната архитектура използват Ино данни от метеорологична мрежа със станции в цяла България [19]. Данните включват измервания от 136 възела на мрежата. Честотата на събиране на данните варира между 60 и 300 секунди. Данните, които са избрани за тестовете са представени в Таблица 7:

Таблица 7. Използвани метеорологични данни

Параметър	Име на променливата	Измервателни единици	Забележка
Сила на сигнала на сензора	Signal	[a.u. %]	Параметър описващ техническото състояние на сензора. Не е включен в анализите.
Съдържание във въздуха на фини прахови частици 2.5 микрона	PM _{2.5}	[ug/m ³]	
Съдържание във въздуха на фини прахови частици 10 микрона	PM ₁₀	[ug/m ³]	PM ₁₀ е производно на PM _{2.5}

Таблица 6 - продължение

Параметър	Име на променливата	Измервателни единици	Забележка
Налягане на въздуха	P	[hPa]	
Температура на въздуха	T	[deg C]	
Относителна влажност на въздуха	RH	[%]	
Космическо лъчение	Gamma	[CPM 5 min]	Бета / гама импулсите са преизчислени в бройки на минута въз основа на докладваните бройки за 5 минути, тъй като това е възприета за докладване величина.
Време на измерването	Unixtime	s	Изминалото време в секунди от 00:00:00 UTC на 1 януари 1970

Ръчно зареждане на данните в Apache Hadoop (HDFS)

Ръчното зареждане на ИНО данните от различни сензори и възли на метеорологичната мрежа в Apache Hadoop HDFS става посредством използване на команди в Linux терминал.

Създаване на външна таблица в Apache Hive / Apache Impala

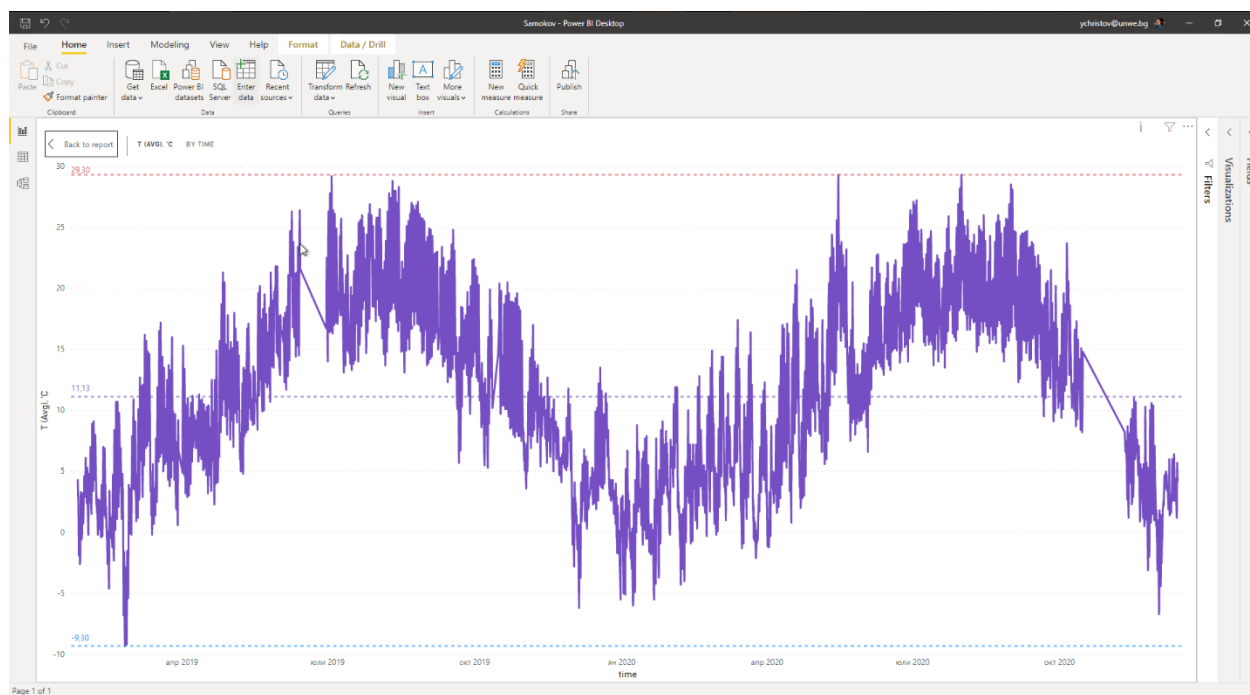
Ръчното създаване на външна Apache Hive таблица с данните от сензорите на от метеорологичната мрежа става посредством SQL заявка изпълнена в Hue редактора (или в beeline за Hive или impala-shell за Apache Impala).

Ръчно зареждане на данните в Microsoft Power BI

Зареждането на данните става посредством използване на ODBC източник в Microsoft Power BI.

Създаване на визуализация със заредените данни

След като ИНО данните от сензорите постъпят в Microsoft Power BI се създава примерна справка и визуализация за изменението температурата на въздуха - Фигура 15.



Фигура 15. Визуализация на метеорологични ИНО данни.

Тестване с транспортни ИНО данни

Източник на транспортни ИНО данни

Източник на транспортните ИНО данни са данните предоставяни от платформата MapQuest. За целите на тестване в настоящия труд са избрани докладите за събитията, пътните инциденти и ремонтните работи по пътната мрежа на Лос Анджелис в САЩ (Таблица 8).

Таблица 8. Използвани транспортни данни.

Параметър	Име на променливата	Измервателни единици	Забележка
Брой събития по пътната мрежа на Лос Анжелис	Incidents	бр.	Включва всички типове регистрирани събития
Текущо време	Unixtime	s	Изминалото време в секунди от 00:00:00 UTC на 1 януари 1970

Ръчно зареждане на данните в Apache Hadoop (HDFS)

Ръчното зареждане на транспортните ИНО данни в CSV формат става с curl команда от Linux конзола.

Създаване на външна таблица в Apache Hive / Apache Impala

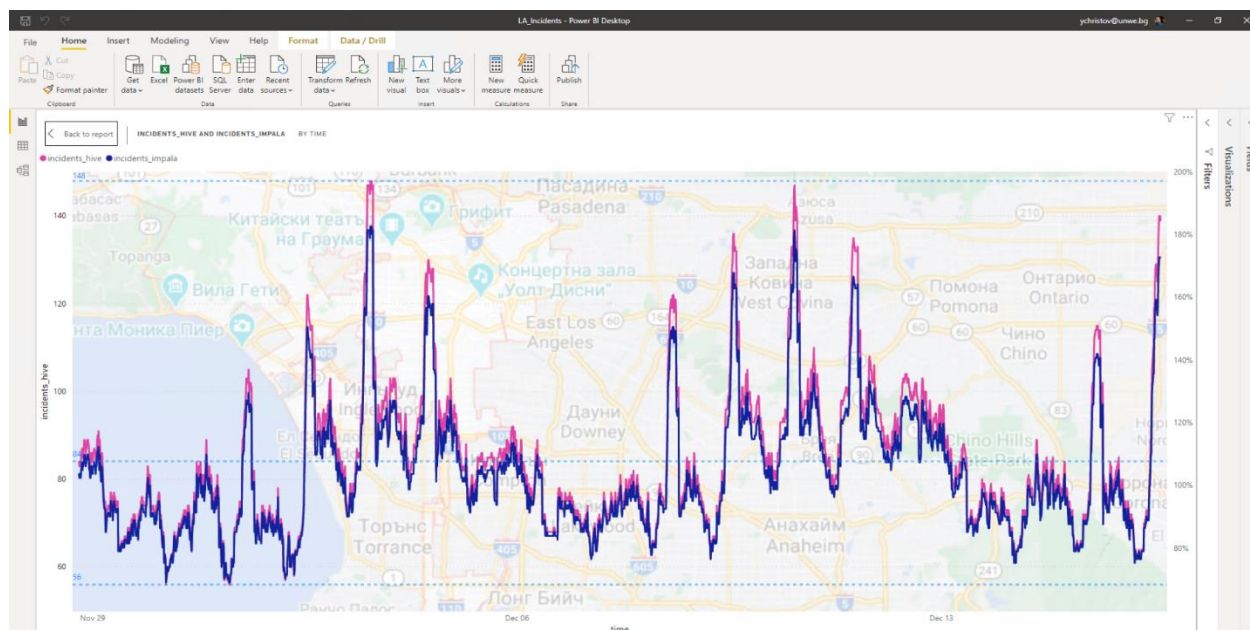
Създаването на външна таблица с данните от инцидентите и ремонтните дейности по пътната мрежа на Лос Анджелис става посредством DDL SQL заявка изпълнена в Hue редактора (или в Apache Hive beeline или Apache Impala impala-shell).

Ръчно зареждане на данните в Microsoft Power BI

Зареждането на данните става посредством използване на ODBC източник в Microsoft Power BI.

Създаване на визуализация със заредените данни

След като ИНО данните за транспортните събития постъпят в Microsoft Power BI се създава примерна справка за броя на инцидентите по пътната мрежа на Лос Анджелис в зависимост от времето. Резултатът е представен на Фигура 16.



Фигура 16. Визуализация на транспортни ИНО данни от Лос Анджелис.

Тестване с финансови ИНО данни

Източник на финансови ИНО данни

Като източник за финансови ИНО данни са използвани част от данните предоставяни от платформата Finnhub. За целите на тестване в настоящата работа са избрани ИНО данните за търговията на финансовите пазари на акциите на три големи компании, а именно Cloudera, Google и Microsoft (Таблица 9).

Таблица 9. Използвани финансови данни.

Параметър	Име на променливата	Измервателни единици	Забележка
Текуща цена	Current	Съответната валута	
Най-висока регистрирана цена	High	Съответната валута	
Най-ниска регистрирана цена	Low	Съответната валута	
Цена при отваряне на пазара	Open	Съответната валута	

Таблица 8 - продължение

Параметър	Име на променливата	Измервателни единици	Забележка
Цена при предишното затваряне на пазара	Previous Close	Съответната валута	
Текущо време	Unixtime	s	Изминалото време в секунди от 00:00:00 UTC на 1 януари 1970

Ръчно зареждане на данните в Apache Hadoop (HDFS)

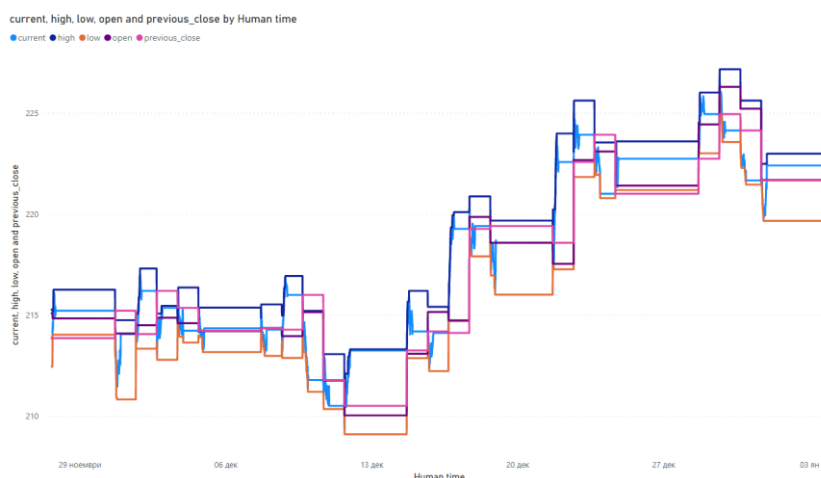
Ръчното зареждане на данните става с curl команда от Linux конзола.

Създаване на външна таблица в Hive / Impala

Създаването на външни таблица с данните от стойностите на акциите на компаниите Cloudera, Google и Microsoft се извършва с аналогични SQL заявки.

Ръчно зареждане на данните в Microsoft Power BI

След като ИНО данните за търговията с акциите на компаниите Cloudera, Google и Microsoft бъдат заредени Microsoft Power BI се създава примерна справка за движението на цените на акциите на съответните компании зависимост от времето. За пример на Фигура 17 е представено движението на акциите на Microsoft.



Фигура 17. Визуализация на финансови ИНО данни за Microsoft.

Тестване на референтната архитектура в автоматичен режим на зареждане на данни

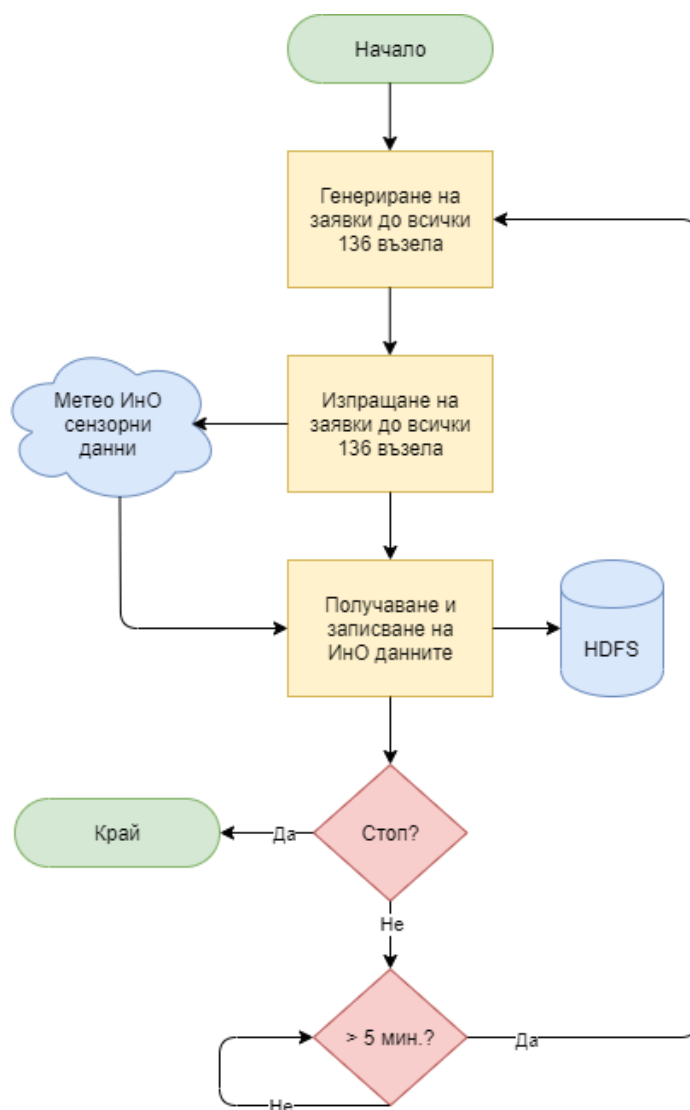
Тестване с метеорологични ИНО данни

Източник на метеорологични ИНО данни

Отново са използвани ИНО данни от метеорологична мрежа със станции в цяла България.

Зареждане на метеорологични ИНО данни в Apache HDFS

На Фигура 18 е представена диаграмата на цикъла за събиране на метеорологични ИНО данни.



Фигура 18. Блок-схема за събиране на метеорологични ИНО данни.

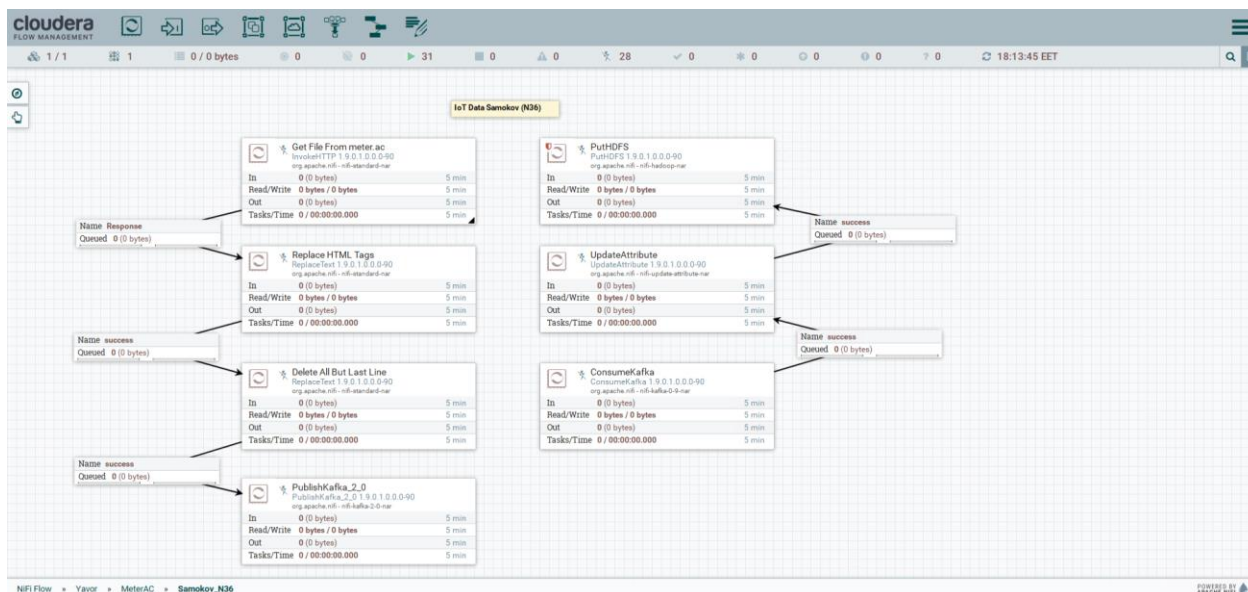
Създаване на външни таблици за метеорологични Ино данни

За да бъдат зареждани автоматично метеорологични Ино данните трябва първо да бъдат създадени външните таблици за избраните възли метеорологичната мрежа. Това е направено чрез два скрипта:

- BASH скрипт за генериране на SQL скрипт за създаване на всички външни таблици;
- SQL скрипт за създаване на таблиците.

Циклично обновяване на метеорологични Ино данни

След приключването на изпълнението на SQL скрипта се пристъпва към зареждането на метеорологични Ино данните посредством създадената референтна архитектура. Създават се потоци за текущите метеорологични Ино данни в Apache NiFi. Детайлите за един примерен поток са представени на Фигура 19.



Фигура 19. Поток на метеорологични Ино данни.

Цикълът се повтаря на всеки 5 минути. В случай на неуспех на някой от процесорите данни не се подават към Apache Kafka и съответно Apache HDFS и Apache Hive. Това става при следващите итерации при успех при получаване на данните от Ино сензорите на метеорологичната мрежа.

Аналогично са изградени потоците за метеорологични ИнО данни и за останалите избрани възли от системата, 136 на брой, за различните метеорологични станции в страната

Добавяне на исторически метеорологични ИнО данни

Без да се спират Apache NiFi потоците за зареждане на текущите метеорологични ИнО данни от избраните възли на мрежата се изпълнява еднократно BASH скрипт за зареждане на историческите данни.

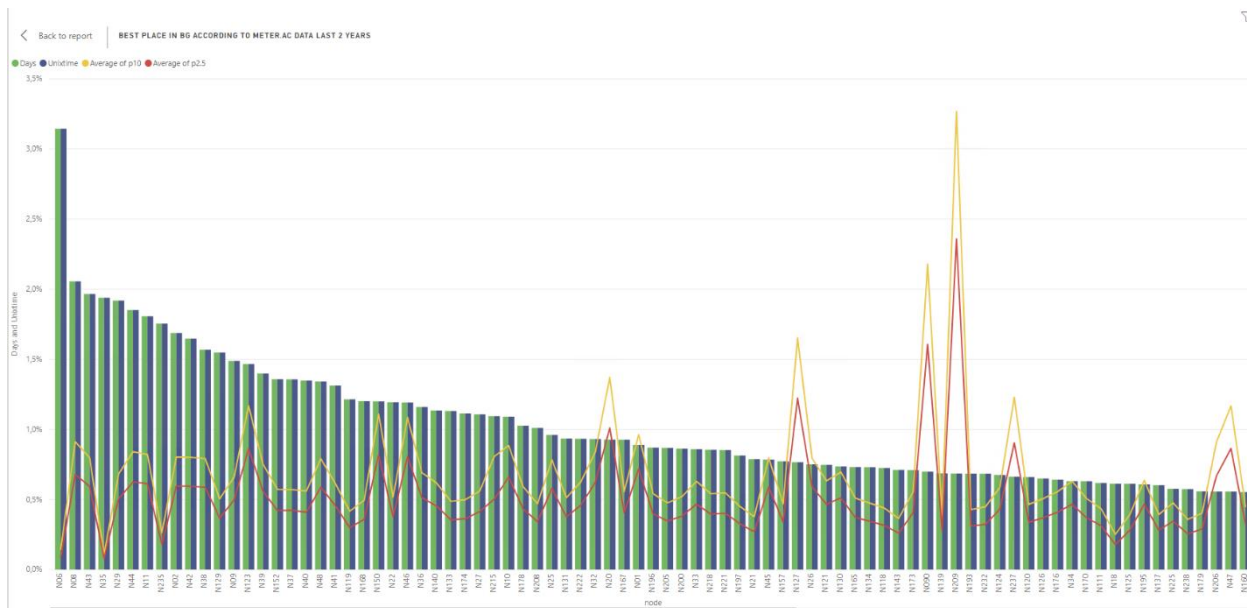
Експериментални резултати

След събирането на метеорологичните ИнО данни от избраните възли на мрежата в България за последните около 2 години беше направен експеримент за откриване на "най-доброто" място за живеене в България съгласно наличните данни. Критериите за избор на "най-добро" място за живеене бяха дефинирани така:

- най-голям брой измервания на относителна влажност на въздуха между 45% и 55% и същевременно температура на въздуха в интервала между 20 °C и 25 °C;
- най-ниски измерени стойности на фини прахови частици.

След обработването на всички събрани ИнО данни (възлизаци на **15 023 776** записа) беше конструирана таблица с **636 446** записа отговарящи на критериите. Интересно е да се отбележи, че референтната архитектура обновява таблицата с новопостъпили данни за по-малко от секунда чрез `impala-shell` и за няколко секунди в Microsoft Power BI. Този факт показва възможностите на клъстера, където е реализирана референтна архитектура за големи ИнО данни.

В Microsoft Power BI беше създадена визуализация на получените данни. Тази визуализация е представена на Фигура 20.



Фигура 20. "Най-добро" място за живот в България.

Отчетливо се откроява два възела от метеорологичната мрежа, а именно:

- възел N06, като "най-добро" място за живот България;
- възел N209, като "най-проблемно" място за живот в България.

Получените резултати за възлите N06 и N209 бяха дискутирани с екипа отговарящ за метеорологичните станции и сензорите в тях и беше установено, че датчиците на възел N06, който надминава всички останали с близо 30% по зададените критерии, са разположени в съвършено помещение с контрол на влажността и температурата на въздуха. За възел N209, при който се наблюдава рязък пик на нивата на фини прахови частици, се оказа, че сензорът за фини прахови частици има хардуерен проблем и отчита неточни и изключително завишени стойности на замърсяване.

Тези резултати са едно добро доказателство за приложимостта на създадената референтна архитектура за анализ и откриване на зависимости, които не са тривиални като се вземе предвид изключително големия брой данни на сензорите за период от 2 години.

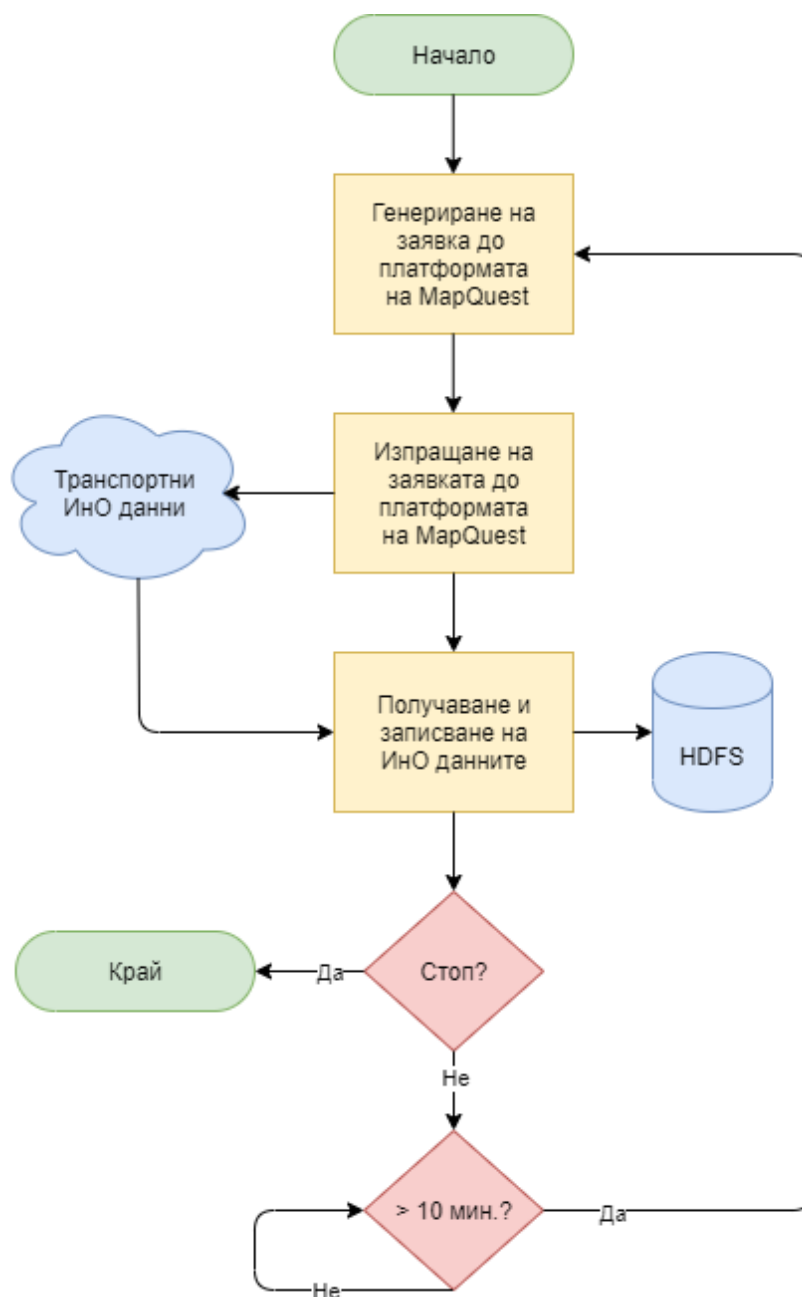
Тестване с транспортни Ино данни

Източник на транспортни Ино данни

Отново са използвани транспортни Ино данни от платформата на MapQuest.

Зареждане на транспортни Ино данни в Apache HDFS

На Фигура 21 е дадена диаграма на цикъла за събиране на транспортни Ино данни.



Фигура 21. Блок-схема за събиране на транспортни Ино данни.

Създаване на HDFS структура за транспортни Ино данни

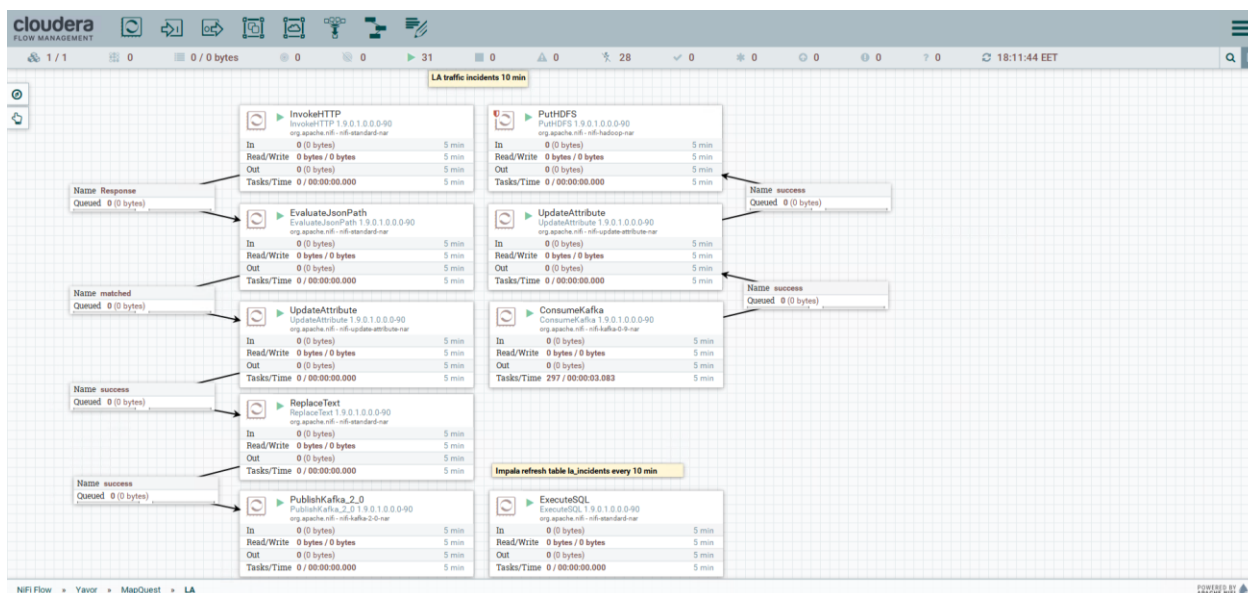
Създаването на необходимата директорийна структура в Apache HDFS става посредством изпълнението на HDFS команди от Linux конзолата.

Създаване на външна таблица за транспортни Ино данни

Създаването на външна таблица с данните от инцидентите по пътната мрежа на Лос Анджелис от платформата на MapQuest става посредством DDL SQL заявка изпълнена в Hue редактора (или в Apache Hive beeline или Apache Impala impala-shell).

Циклично обновяване на транспортните Ино данни

Зареждането на транспортните Ино данни става с помощта на компонентите Apache NiFi, Apache Kafka и Apache Hive (Apache Impala) от създадената референтна архитектура. Създава се поток за текущите транспортни Ино данни в Apache NiFi представен на Фигура 22.



Фигура 22. Поток на транспортни Ино данни за гр. Лос Анджелис.

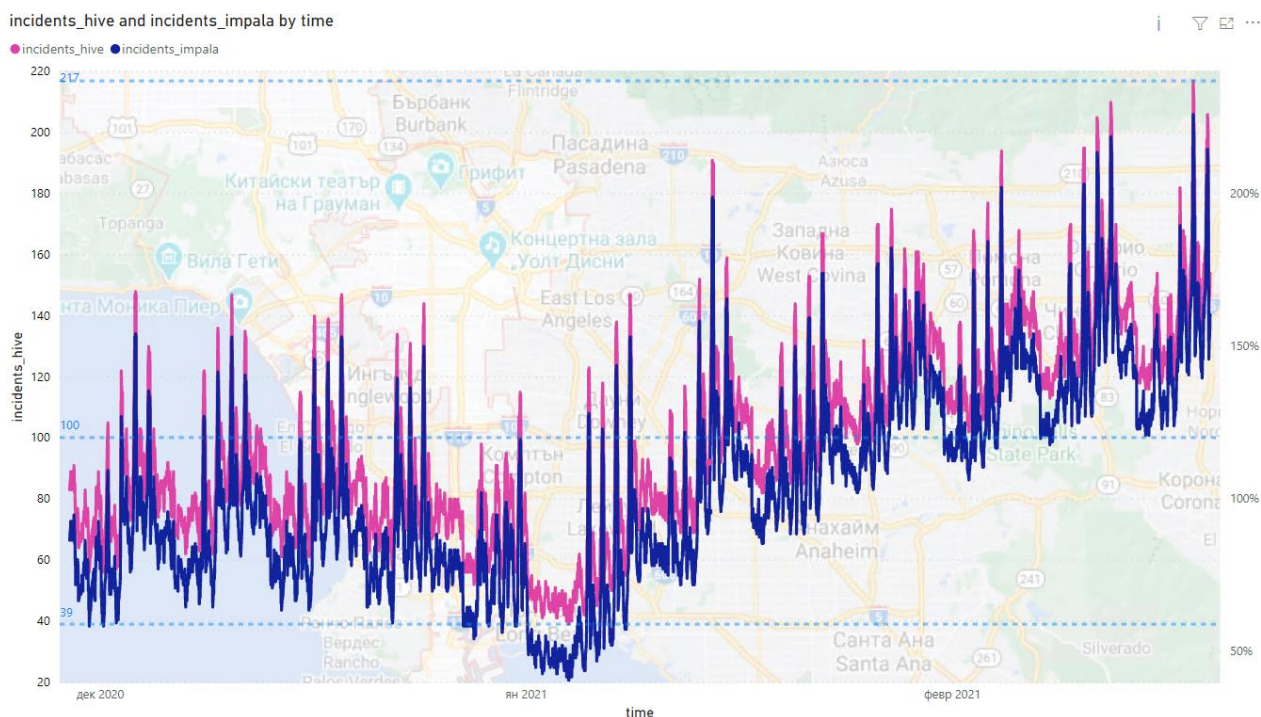
Цикълът се повтаря на всеки 10 минути. В случай на неуспех на някой от процесорите данни не се подават към Apache Kafka и съответно Apache Hadoop HDFS и Apache Hive (Apache Impala). Това става при следващите итерации при

успех при получаване на данните от ИНО платформата MapQest за текущите стойности за инцидентите по пътната мрежа.

Експериментални резултати

Експерименталните данни за транспортните събития по пътищата на Лос Анжелис обхващат период от три месеца (ноември 2020 - февруари 2021).

Беше извършено експериментално сравнение на бързодействието на зареждане на транспортните ИНО данни (повече от 12 500 записа във външната таблица `la_incidents`) в Microsoft Power BI посредством Apache Impala и Apache Hive. Зареждането на данните чрез Apache Impala отнема средно около 5 секунди докато зареждането на същите данни чрез Apache Hive отнема средно около 35 секунди. На Фигура 23 са представени графиките на инцидентите по пътната мрежа в Лос Анжелис съответно получени с Apache Impala (`median`) и Apache Hive (`absolute`)



Фигура 23. Събития на територията на Лос Анжелис.

На графиката на Фигура 23 отчетливо се вижда седмичен цикъл на броя на събитията по пътната мрежа на Лос Анжелис с характерни пикове през почивните дни (петък, събота и неделя). Друга характерна, логично очаквана особеност, е

локалния минимум в дните 31.12.2020 - 03.01.2021 г., когато хората са по домовете си. Получената визуализация и нейните параметри доказват експериментално адекватността на приложената референтна архитектура за обработка и анализ на транспортни ИНО данни.

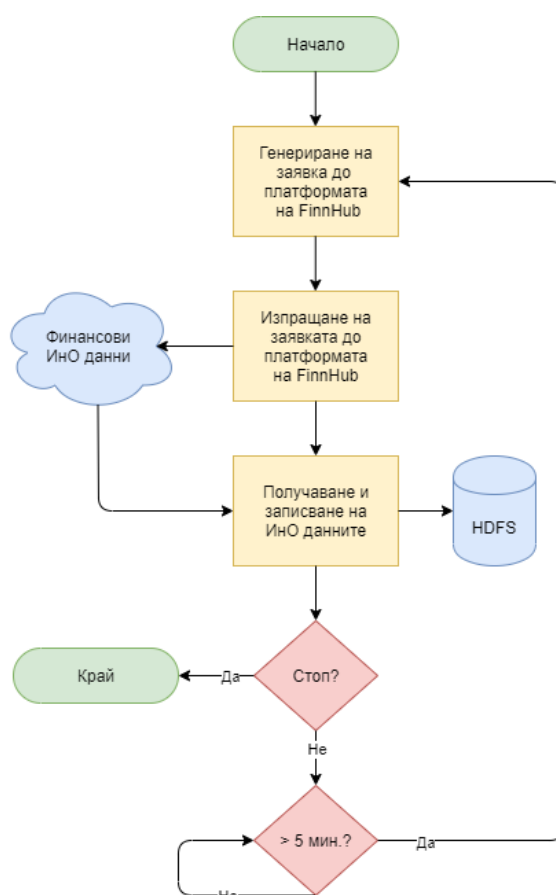
Тестване с финансови ИНО данни

Източник на финансови ИНО данни

За източник за финансови ИНО данни се използва отново платформата на Finnhub. За целите на тестване в настоящата работа са избрани ИНО данните за търговията на финансовите пазари на акциите на Cloudera, Google и Microsoft.

Автоматично зареждане на данни в Apache HDFS

На Фигура 24 е дадена диаграма на цикъла за събиране на финансови ИНО данни.



Фигура 24. Блок-схема за събиране на финансови ИНО данни.

Създаване на HDFS структура за финансови Ино данни

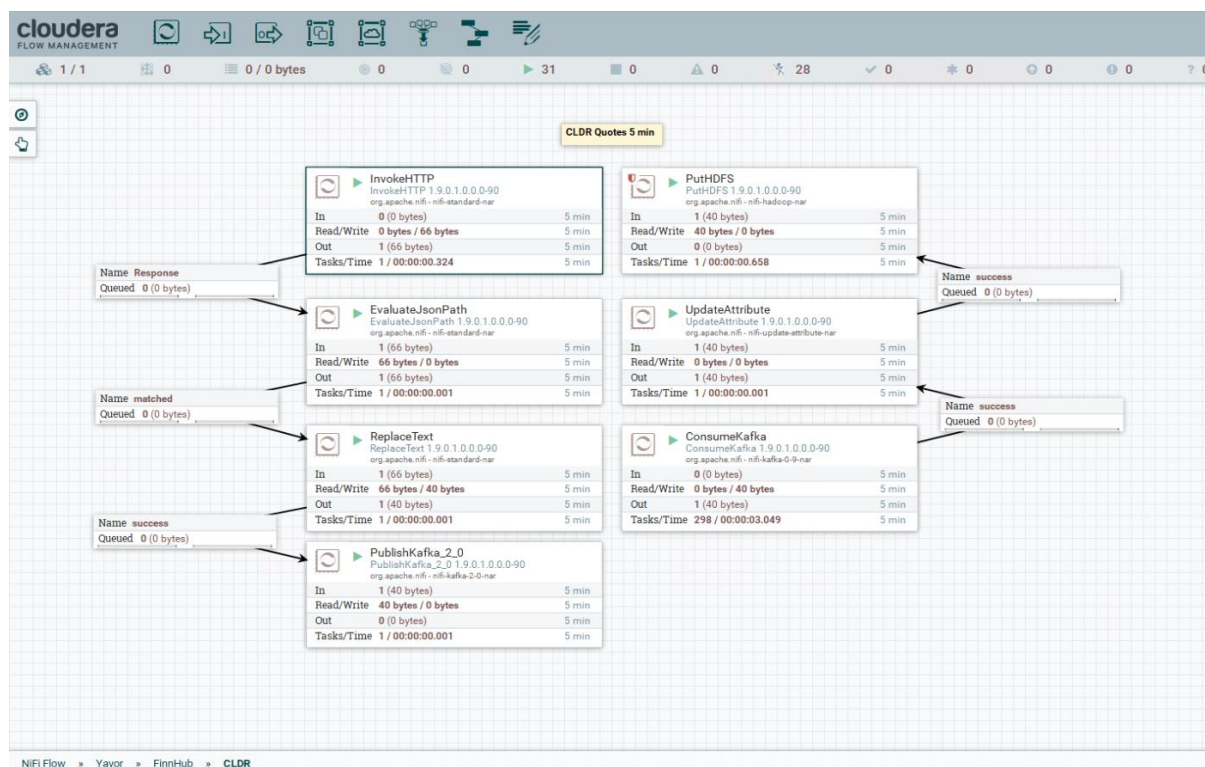
Създаването на необходимата директорийна структура в Apache Hadoop HDFS става посредством изпълнението на HDFS команди от Linux козолата.

Създаване на външни таблици за финансови Ино данни

Създаването на външни таблици с данните за търговията на акциите на компаниите Cloudera, Google и Microsoft става посредством DDL SQL заявки изпълнени в Hue редактора (или в Hive beeline или Impala impala-shell).

Циклично обновяване на финансовите Ино данни

Зареждането на финансовите Ино данни става с помощта на компонентите Apache NiFi, Apache Kafka и Apache Hive (Apache Impala) от създадената референтна архитектура. Създават се потоци за текущите финансовите Ино данни в Apache NiFi. Детайлите за потока за Cloudera са представени на Фигура 25. Потоците за финансовите данни от Google и Microsoft са аналогични.



Фигура 25. Поток на финансови Ино данни за акциите на Cloudera.

Цикълът се повтаря на всеки 5 минути. В случай на неуспех на някой от процесорите данни не се подават към Apache Kafka и съответно Apache Hadoop HDFS и Apache Hive (Apache Impala). Това става при следващите итерации при успех при получаване на данните от ИНО платформата FinnHub за текущите стойности за акциите на съответната компания.

Експериментални резултати

Експерименталните данни за движението на акциите на компаниите Cloudera (CLDR), Google (GOOG) и Microsoft (MSFT) обхващат период от три месеца (ноември 2020 - февруари 2021) с по над 25 000 записа за всяка компания.

На базата на резултатите от извършено експериментално сравнение на бързодействието на зареждане на ИНО данни в Microsoft Power BI посредством Apache Impala и Apache Hive беше решено данните за движението на акциите да се зареждат само с Apache Impala. На Фигура 26 е представена графиката за движението на акциите на Microsoft. Графиките за движенията на акциите на Google и Cloudera изглеждат аналогично.



Фигура 26. Движение на акциите на Microsoft - ноем. 20 г. - февр.. 21 г.

На графиката на Фигура 26 отчетливо можем да забележим седмичен цикъл на движението на акциите на Microsoft до началото на януари 2021 г. След 4.01.2021 г. платформата FinnHub промени условията за предоставяне на данни за разработчици. След тази дата акциите на компанията са представени с единични стойности

за всеки ден. Това ясно личи на представената графиката. С други думи, предложената референтна архитектура се адаптира към промени в условията на постъпване на Ино данни без необходимост от странична намеса. Получената визуализация и нейните параметри доказват експериментално адекватността на приложената референтна архитектура за обработка и анализ на финансови Ино данни.

Заключение

Създадената компонентна референтна архитектура е мощен инструмент за обработка от край до край на големи данни от хетерогенни Ино източници. Тестовите в ръчен и автоматичен режим на зареждане на данните доказват:

- гъвкавост на референтната архитектура спрямо източниците на големи Ино данни;
- адекватност на референтната архитектура спрямо реалните процеси протичащи в системите генериращи големи Ино данни;
- способност за откриване на нетривиални изключения в набора от големи Ино данни;
- стабилност на референтната архитектура при настъпване на промени в системите генериращи големи Ино данни;
- възможност за реализация на референтната архитектура в различни мащаби (малки и големи клъстери);
- автоматичният режим на работа е по-ефективния от двата режима и води до значимо повишаване на възможностите и удобството при работа със създадената архитектура.

6. ЗАКЛЮЧЕНИЕ

В началото на настоящия дисертационен труд бяха дефинирани три основни работни хипотези, а именно:

Хипотеза I:

Възможно е да се създаде референтна архитектура, приложима за обработка и работа с данни от Интернет на обектите.

Хипотеза II:

Възможно е да се даде количествена оценка за референтна архитектура, като се създаде методология за оценка по различни показатели.

Хипотеза III:

Възможно е да се постигне ефективна интеграция на компонентите на референтната архитектура, за работа с големи данни.

С оглед на представеното дотук в настоящия дисертационен труд можем да заключим, че и трите хипотези бяха доказани успешно.

7. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Christov Y., Cleaning data with cloud based platform toolsets, 5th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2017), November 3-4th, 2017, UNWE, Sofia, Bulgaria, ISSN 2367-7635 (Print), ISSN 2367-7643 (online), pp. 503-509.

2. Christov Y., Big Data Sets Modeling with Differential Neural Networks, Новите информационни технологии и големите данни: възможности и перспективи при анализите и управленските решения в бизнеса, икономиката и социалната сфера, Сборник с доклади от международна научна конференция организирана от катедра „Статистика и иконометрия - със съдействието на ф-т „Приложна информатика и статистика - и в партньорство с Националния статистически институт, 6 – 7 юни 2019, издателски комплекс УНСС, ISBN 978-619-232-292-2, стр. 287-292.

3. Christov Y., Corporate Big Data and Hybrid Integration Platforms, Proceedings of the International Conference Information and communication and technologies in business and education dedicated to the 50th anniversary of the Department of Informatics, October 19th 2019, University of Economics – Varna, Varna, Bulgaria, ISBN 978-954-21-1004-0, pp. 306-313.

4. Christov Y., Building Personal Virtual Private Networks in Public Cloud Platforms, International scientific conference Technics. Technologies. Education. Safety., Proceedings year IV, issue 1 (8), volume 1, Technics and technology. Information technology, June 08-11th 2020, Borovets, Bulgaria, ISSN 2535-0315 (Print), ISSN 2535-0323 (online), pp. 33-34.

5. Boyanov L., Kisimov V., Christov Y., Evaluating IoT Reference Architecture, 2020 International Conference Automatics and Informatics (ICAI), October 1-3rd 2020, Varna, Bulgaria, IEEE Xplore, January 7th 2021, Electronic ISBN:978-1-7281-9308-3, (PoD) ISBN:978-1-7281-9309-0, DOI: 10.1109/ICAI50593.2020.9311357, Publisher: IEEE

ЛИТЕРАТУРА

- [1] J. Manyika и съавт., „The Internet of Things: Mapping the Value Beyond the Hype“. McKinsey Global Institute (MGI), Юни 2015, Отворен на: Фев 13, 2021. [Онлайн]. Достъпен на: https://www.mckinsey.com/~media/McKinsey/Industries/Technology%20Media%20and%20Telecommunications/High%20Tech/Our%20Insights/The%20Internet%20of%20Things%20The%20value%20of%20digitizing%20the%20physical%20world/Unlocking_the_potential_of_the_Internet_of_Things_Executive_summary.pdf.
- [2] K. Ashton, „That ‘Internet of Things’ Thing“. RFID JOURNAL, Юни 22, 2009, Отворен на: Фев 12, 2021. [Онлайн]. Достъпен на: <https://www.rfidjournal.com/that-internet-of-things-thing>.
- [3] D. Evans, „The Internet of Things How the Next Evolution of the Internet Is Changing Everything“. Cisco, Април 2011, Отворен на: Фев 13, 2021. [Онлайн]. Достъпен на: https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf.
- [4] T. A. Dang, „Big Data: Lambda Architecture in a nutshell“, *Medium*, Октомври 20, 2020. <https://levelup.gitconnected.com/big-data-lambda-architecture-in-a-nutshell-fd5e04b12acc> (отворен на Фев 19, 2021).
- [5] M. Mallappallil, J. Sabu, A. Gruessner, и M. Salifu, „A review of big data and medical research“, *SAGE Open Med*, том 8, Юни 2020, doi: 10.1177/2050312120934839.
- [6] UNWE, „Научни трудове на УНСС - Том 1/2019“, Юни 11, 2019. <http://unwe-research-papers.org/bg/journalissues/list/135> (отворен на Фев 19, 2021).
- [7] „ISO/IEC/IEEE 42010:2011(en), Systems and software engineering — Architecture description“. <https://www.iso.org/obp/ui/#iso:std:iso-iec-ieee:42010:ed-1:vi:en> (отворен на Фев 13, 2021).
- [8] Y. Christov, „Cleaning data with cloud based platform toolsets“, в *Proceedings Of The 7th International Conference On Application Of Information And Communication Technology And Statistics In Economy And Education Icaictsee-2017*, Sofia, 2017, стр. 503–509.
- [9] „Apache NiFi“. <https://nifi.apache.org/> (отворен на Фев 15, 2021).
- [10] „Apache Kafka“. <https://kafka.apache.org/> (отворен на Фев 15, 2021).
- [11] „Apache Hadoop“. <http://hadoop.apache.org/> (отворен на Фев 15, 2021).
- [12] „Apache Hive“. <http://hive.apache.org/> (отворен на Фев 15, 2021).
- [13] „Apache Impala“. <https://impala.apache.org/> (отворен на Фев 15, 2021).
- [14] „Hue - The open source SQL Assistant for Data Warehouses“. <https://gethue.com/new-sql-editor/> (отворен на Фев 15, 2021).
- [15] „Microsoft Power BI“. <https://powerbi.microsoft.com/en-us/> (отворен на Фев 15, 2021).
- [16] „ISO/IEC 30141:2018“, *ISO*. <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65695.html> (отворен на Фев 13, 2021).

- [17] L. Boyanov, V. Kisimov, и Y. Christov, „Evaluating IoT Reference Architecture“, в *2020 International Conference Automatics and Informatics (ICAI)*, Окт 2020, стр. 1–5, doi: 10.1109/ICAI50593.2020.9311357.
- [18] „Cloud Market Share – a Look at the Cloud Ecosystem in 2021“, *Kinsta*, Ное 11, 2019. <https://kinsta.com/blog/cloud-market-share/> (отворен на Фев 19, 2021).
- [19] A. Terziyski, S. Tenev, V. Jeliazkov, N. Jeliazkova, и N. Kochev, „METER.AC: Live Open Access Atmospheric Monitoring Data for Bulgaria with High Spatiotemporal Resolution“, *Data*, том 5, бр. 2, Art. бр. 2, Юни 2020, doi: 10.3390/data5020036.

СПИСЪК НА ФИГУРИТЕ

Фигура 1. Структура и връзки на основните компоненти на ИноО.	12
Фигура 2. Ламбда архитектура.	13
Фигура 3. Основни показатели на големите данни (V^4).	15
Фигура 4. Принцилна схема на ИноО РА.	26
Фигура 5. Архитектура на Apache NiFi.	27
Фигура 6. Архитектура на Apache Kafka.	28
Фигура 7. Архитектура на Apache Hadoop.	29
Фигура 8. Архитектура на Apache Hive.	30
Фигура 9. Архитектура на Apache Impala.	31
Фигура 10. Ние компоненти и типични приложения.	32
Фигура 11. Визуализация с Microsoft Power BI. [15]	33
Фигура 12. Графично представяне на оценките на РА.	37
Фигура 13. Оценка на ИноО РА за Microsoft, Google и Cloudera.	42
Фигура 14. Принцилна схема на предложената компонентна ИноО РА.	45
Фигура 15. Визуализация на метеорологични ИноО данни.	48
Фигура 16. Визуализация на транспортни ИноО данни от Лос Анджелис. ...	50
Фигура 17. Визуализация на финансови ИноО данни за Microsoft.	51
Фигура 18. Блок-схема за събиране на метеорологични ИноО данни.	52
Фигура 19. Поток на метеорологични ИноО данни.	53
Фигура 20. "Най-добро" място за живот в България.	55
Фигура 21. Блок-схема за събиране на транспортни ИноО данни.	56
Фигура 22. Поток на транспортни ИноО данни за гр. Лос Анджелис.	57
Фигура 23. Събития на територията на Лос Анджелис.	58
Фигура 24. Блок-схема за събиране на финансови ИноО данни.	59
Фигура 25. Поток на финансови ИноО данни за акциите на Cloudera.	60
Фигура 26. Движение на акциите на Microsoft - ноем. 20 г. - февр. 21 г.	61

СПИСЪК НА ТАБЛИЦИТЕ

Таблица 1. Типове големи данни.	17
Таблица 1 - продължение.....	18
Таблица 2. Класификация на източниците на Ино данни.	21
Таблица 3. Количествена оценка на референтни архитектури за Ино.	35
Таблица 4. Оценка на конкретни референтни архитектури за Ино.....	38
Таблица 5. Количествена оценка на предложената архитектура.	43
Таблица 6. Използвани метеорологични данни.....	46
Таблица 7. Използвани транспортни данни.....	49
Таблица 8. Използвани финансови данни.....	50