НТУУ «КПІ»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА СИСТЕМНОГО ПРОЕКТУВАННЯ

Презентація до дипломної роботи на тему: “Розподілена обробка великих масивів даних” на прикладі онлайн ринків

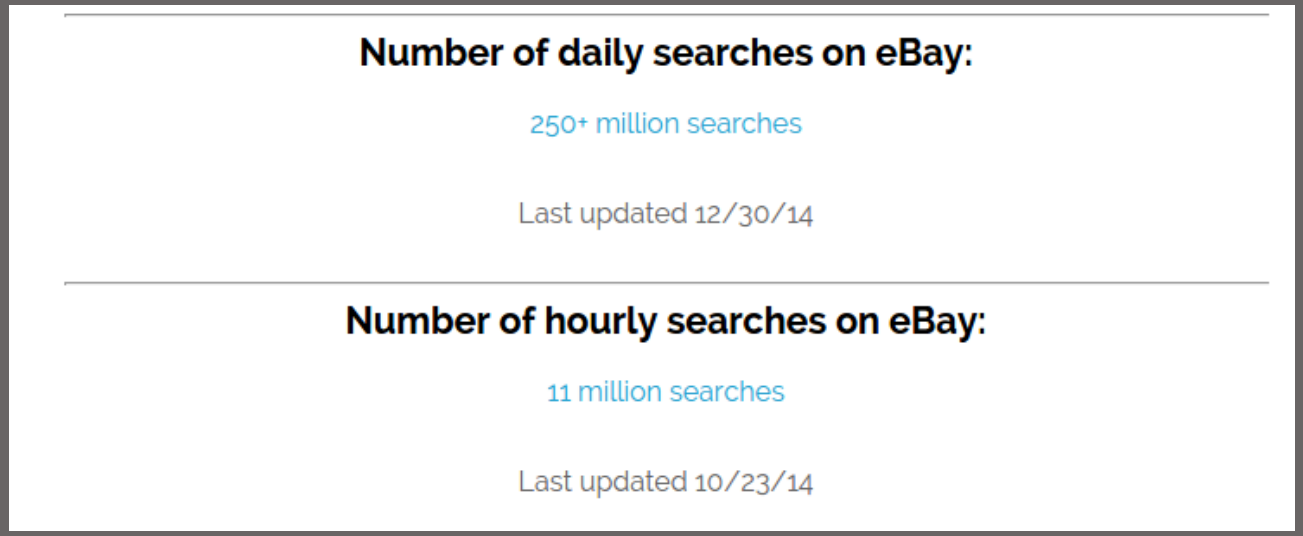
Виконав: Варга І. Ю.

Керівник: Корначевський Я.І.

Київ 2015

Актуальність

Сучасна розповсюдженість ринків:



~ 22000 записів в секунду

Aliexpress являється найбільш багаточисельним серед ринків середньому ми маємо 600 мільйонів користувачів в місяць

Зосередивши увагу на даних по користувачах по покупкам і по переглядам товарів можна вивести певну статистику й на її основі зробити статистику та прогноз

Методи аналізу

Методи аналізу великих данних :

- Краудсорсінг
- Machine learning
- Прогнозна аналітика
- Статистичний анализ
- Візуалізація аналітичних даних

Регресія

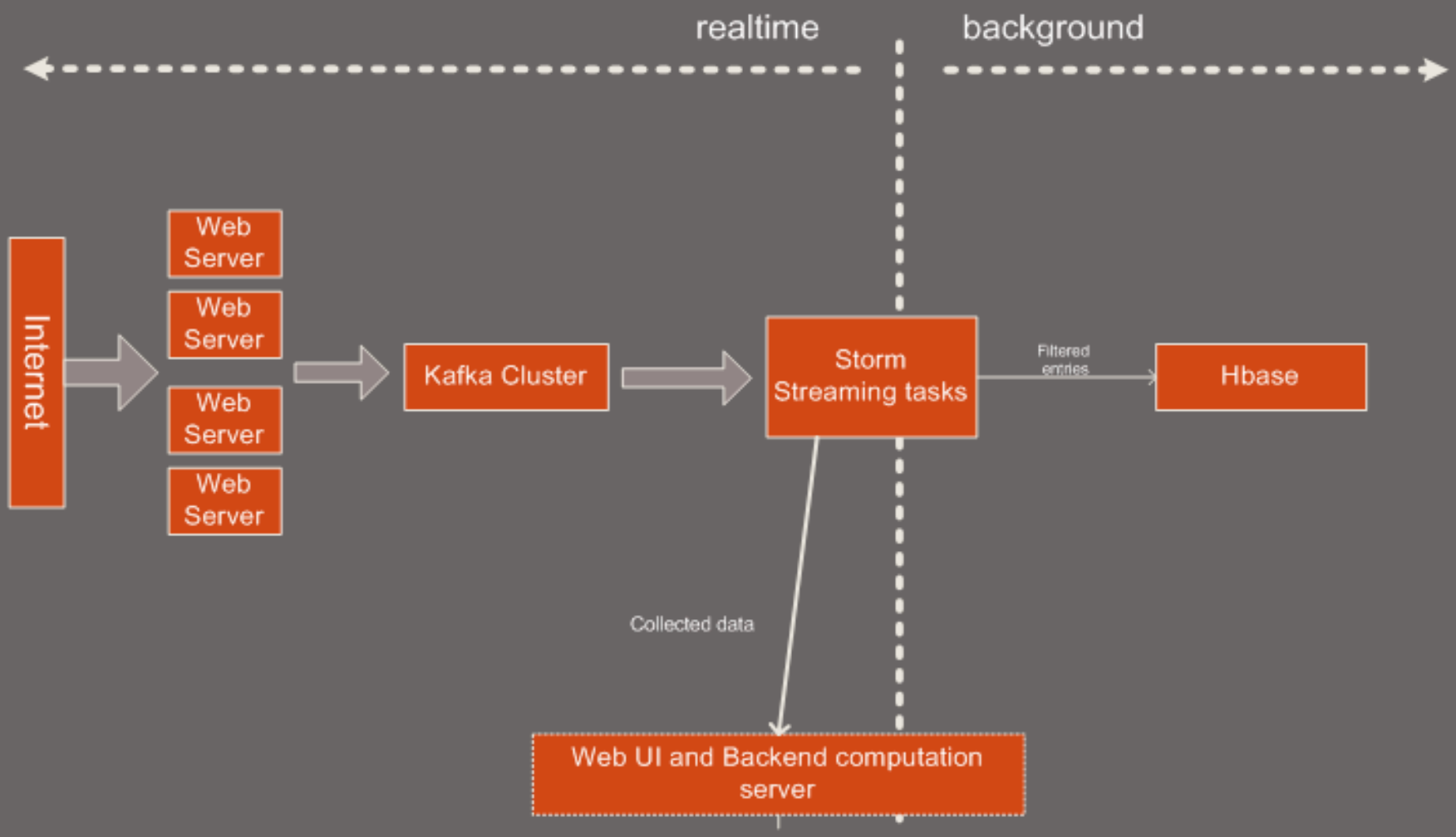
Лінійна регресія (англ. Linear regression) - використовувана в статистиці регресійна модель залежності однієї (що пояснюється, залежної) змінної y від іншої або кількох інших змінних (факторів, регресорів, незалежних змінних) x з лінійною функцією залежності.

В нашому випадку не було акценту на самі методи прогнозування, тому була вибрана модель лінійної регресії. Як виявилось вона доволі вдало справляється зі своїми обов’язками

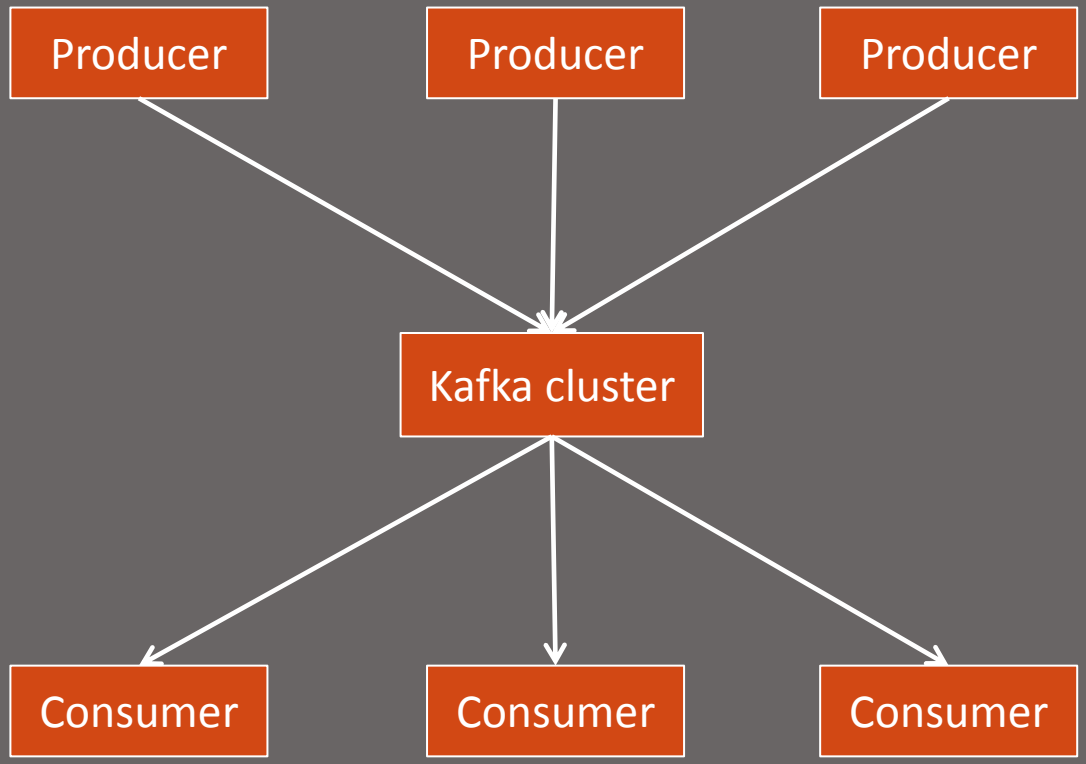
Вибір і порівняння дистрибутивів

	Hortonworks	Cloudera	MapR
Performance and Scalability			
Data Ingest	Batch	Batch	Batch and streaming writes
Metadata Architecture	Centralized	Centralized	Distributed
HBase Performance	Latency spikes	Latency spikes	Consistent low latency
NoSQL Applications	Mainly batch applications	Mainly batch applications	Batch and online/real-time applications
Dependability			
High Availability	Single failure recovery	Single failure recovery	Self healing across multiple failures
MapReduce HA	Restart jobs	Restart jobs	Continuous without restart
Upgrading	Planned downtime	Rolling upgrades	Rolling upgrades
Replication	Data	Data	Data + metadata
Snapshots	Consistent only for closed files	Consistent only for closed files	Point-in-time consistency for all files and tables
Disaster Recovery	No	File copy scheduling (BDR)	Mirroring
Manageability			
Management Tools	Ambari	Cloudera Manager	MapR Control System
Volume Support	No	No	Yes
Heat map, Alarms, Alerts	Yes	Yes	Yes
Integration with REST API	Yes	Yes	Yes
Data and Job Placement Control	No	No	Yes
Data Access			
File System Access	HDFS, read-only NFS	HDFS, read-only NFS	HDFS, read/write NFS (POSIX)
File I/O	Append only	Append only	Read/write
Security: ACLs	Yes	Yes	Yes
Wire-level Authentication	Kerberos	Kerberos	Kerberos, Native

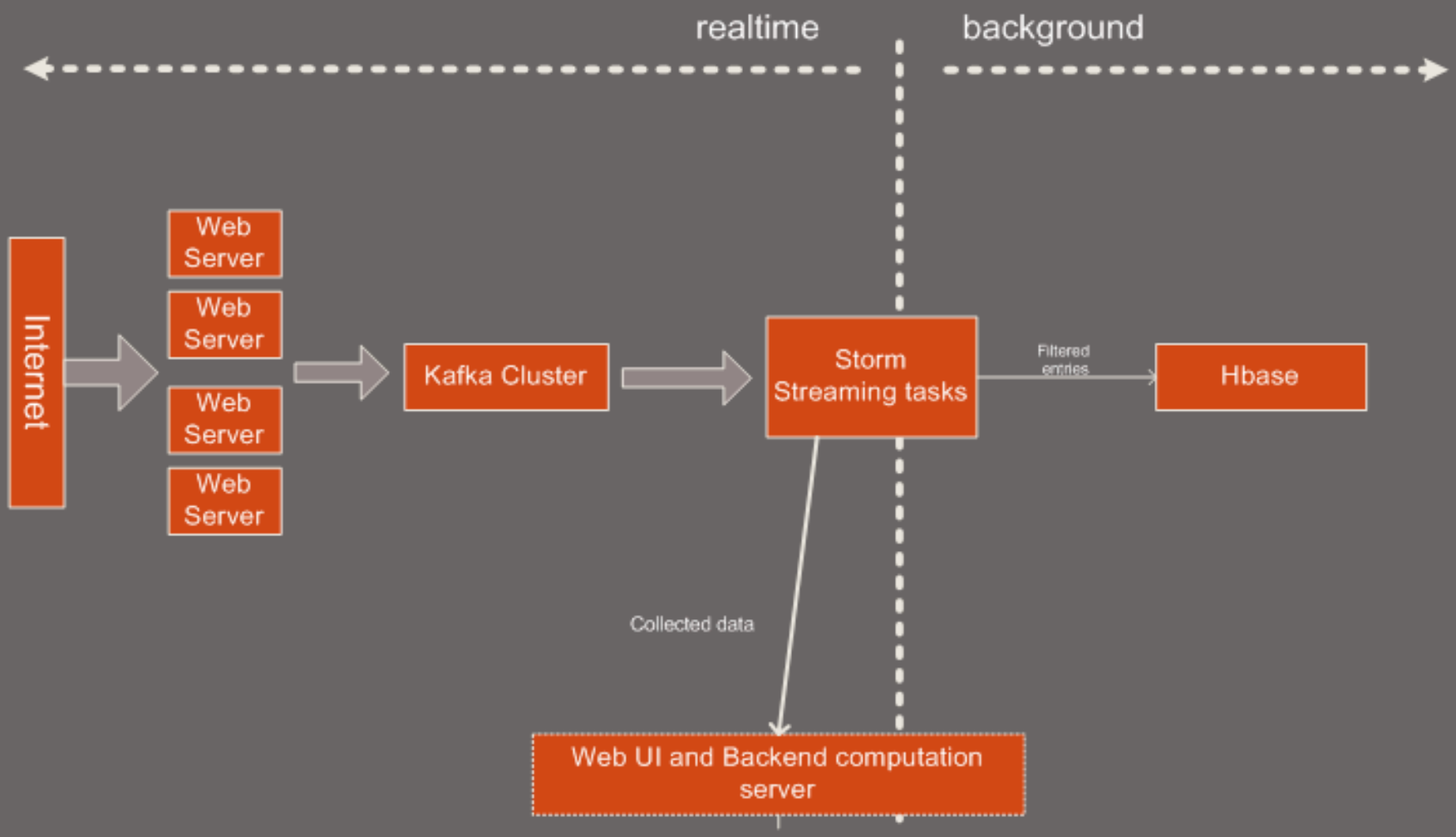
LAMBDA ARCHITECTURE



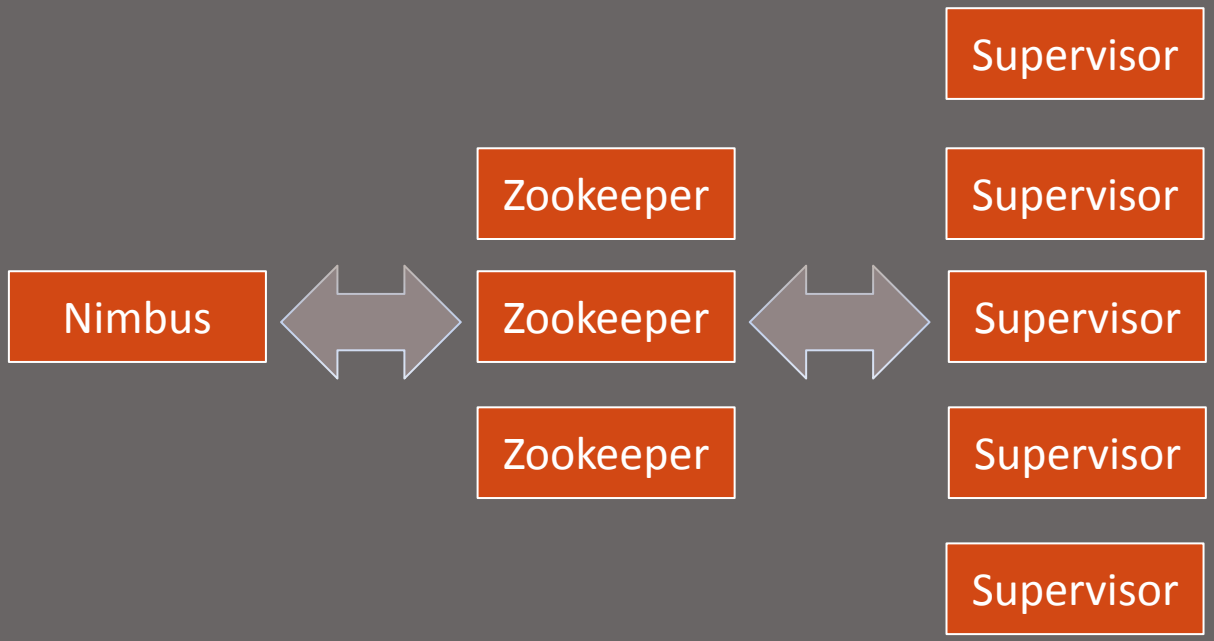
Kafka Cluster



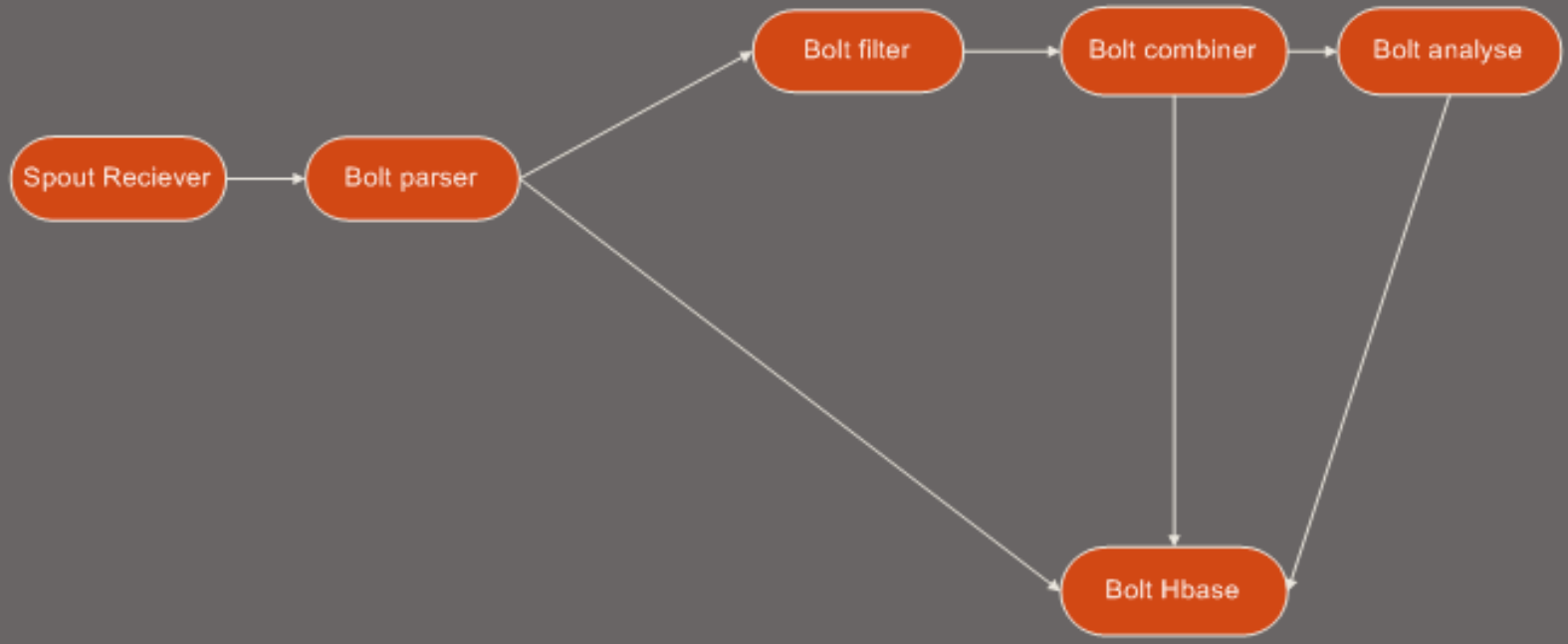
LAMBDA ARCHITECTURE



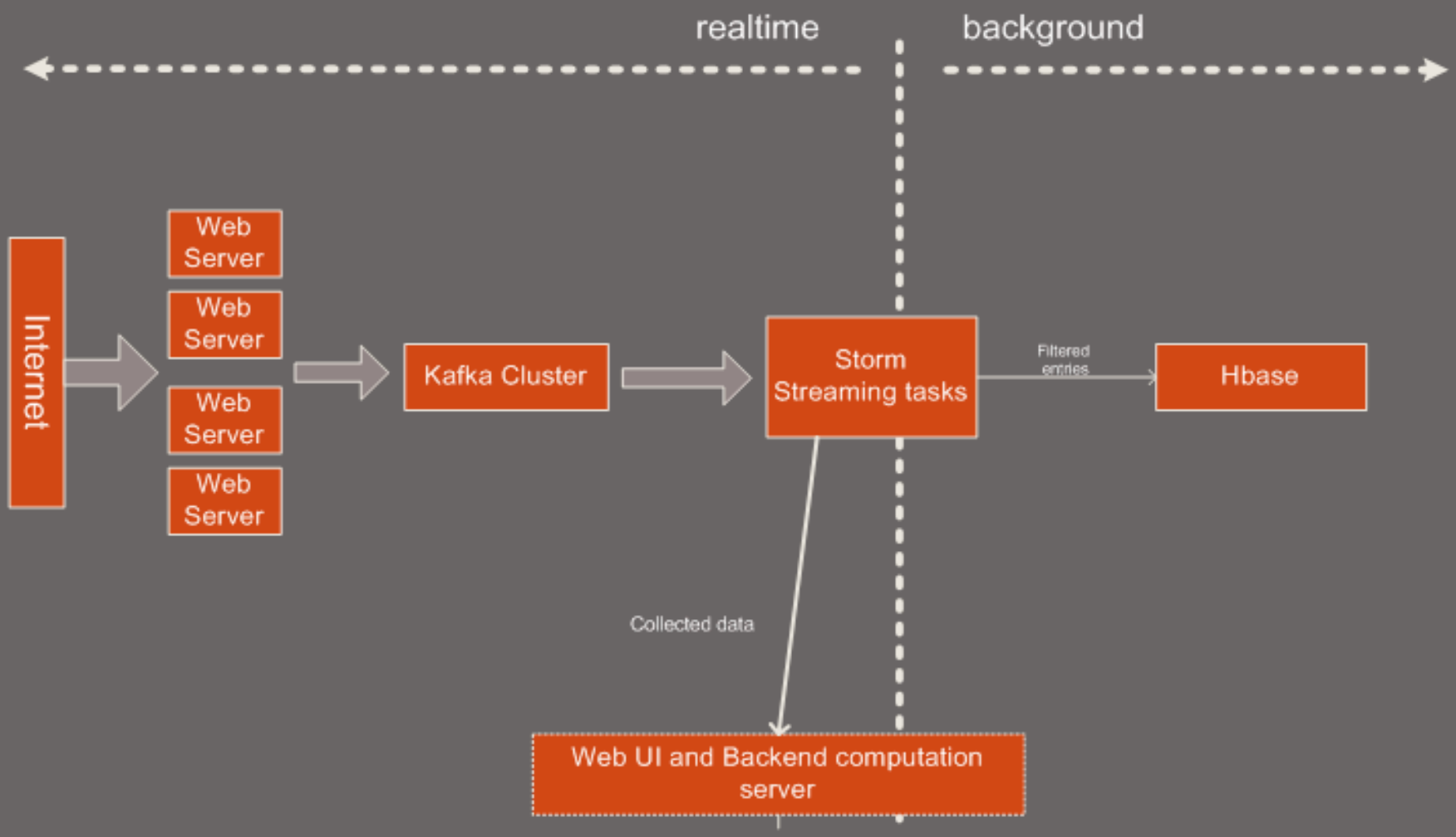
Basic Storm Cluster



Internal topology architecture



LAMBDA ARCHITECTURE



System Requirements

30 cluster nodes

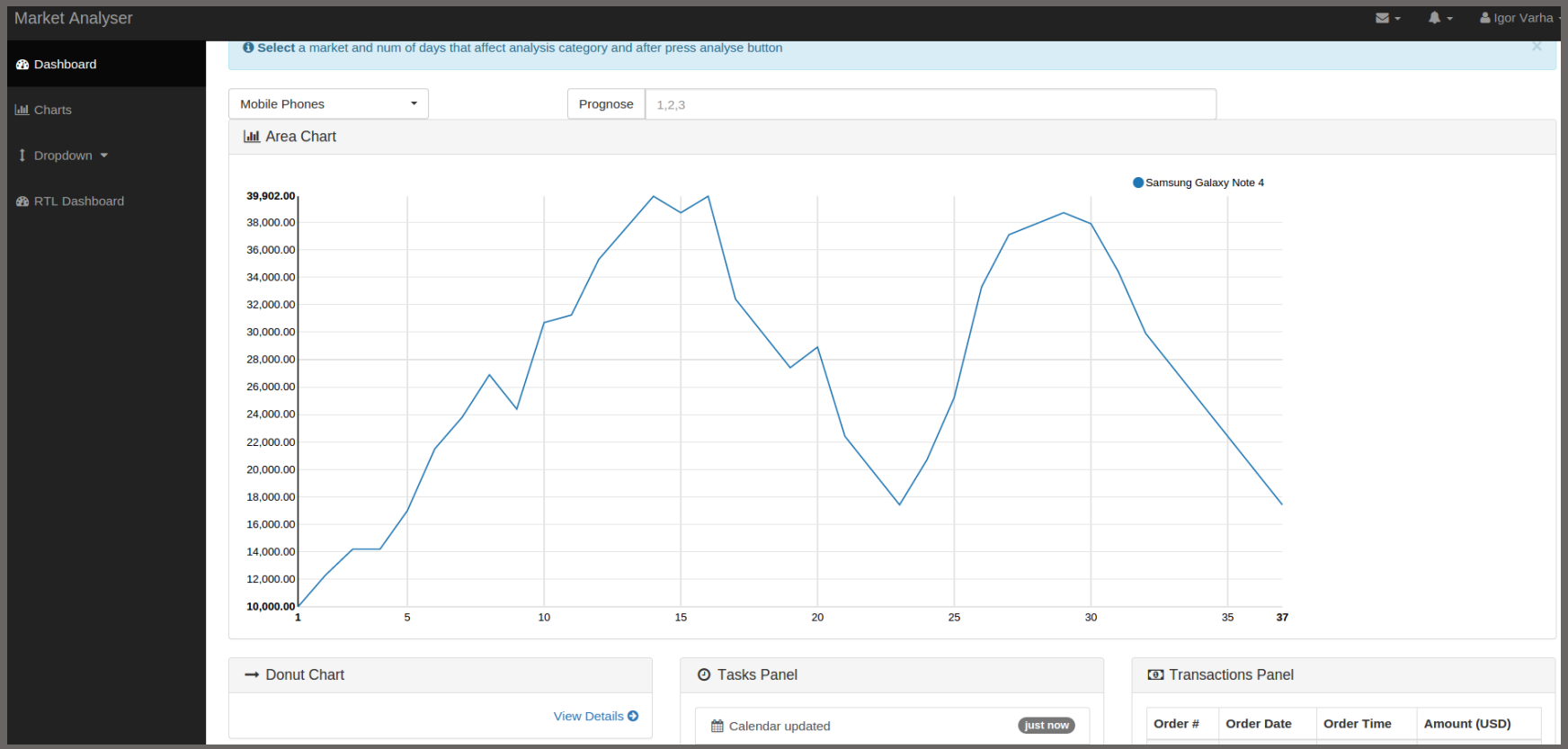
~900 queries/sec

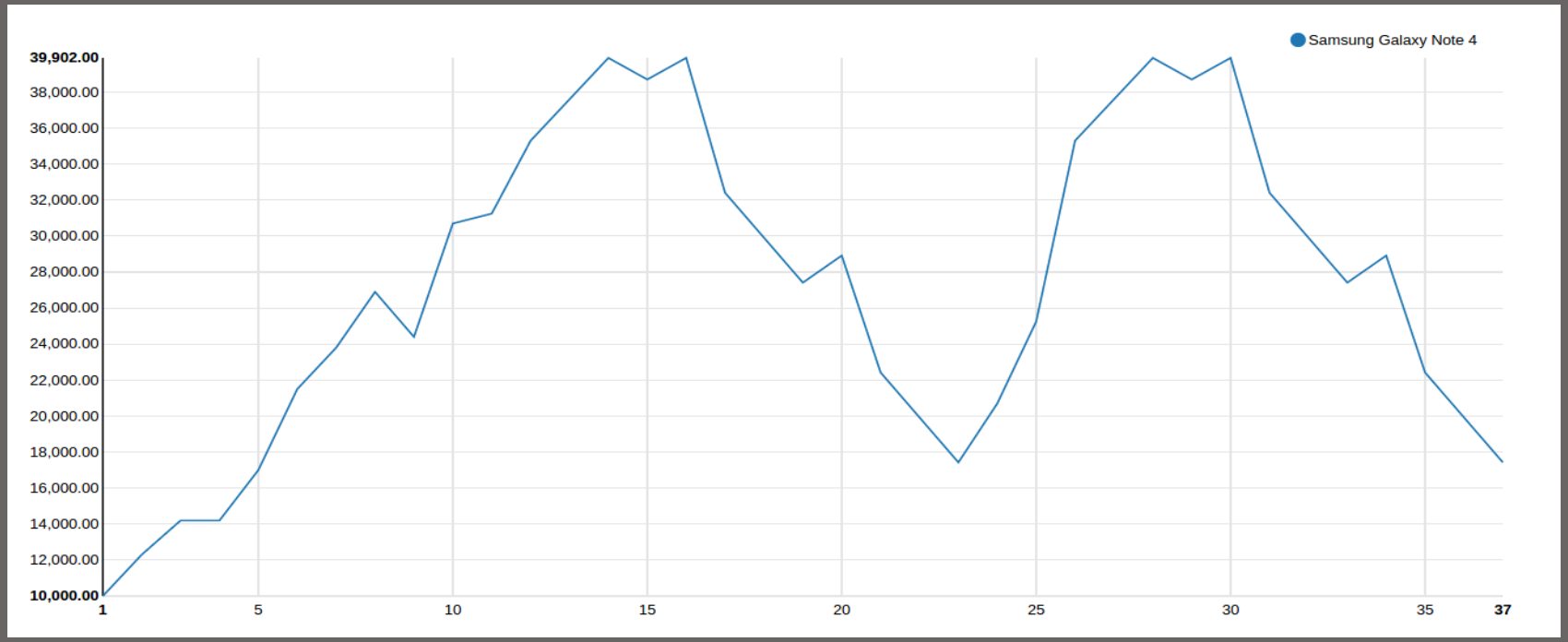
500TiB storage

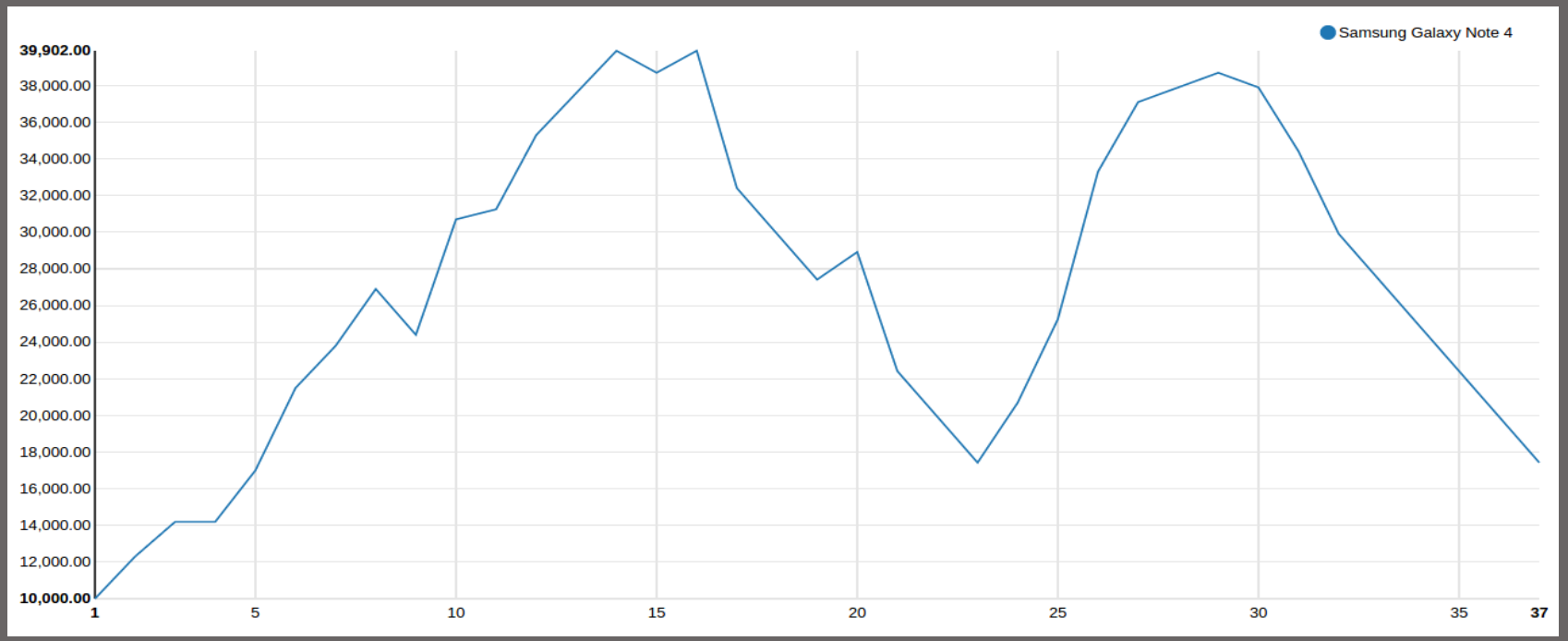
2 PiB storage in future

Unit	Type
CPU	64-bit
OS	Red Hat, CentOS, SUSE, or Ubuntu
Memory	8 GB minimum
Disk	Raw, unformatted drives and partitions
DNS	Hostname, reaches all other nodes
Users	Common users across all nodes; passwordless ssh (optional)
Java	Must run Java
Other	NTP, Syslog, PAM

Результати


Market Analyser
Igor Varha
Dashboard
Charts
Dropdown
RTL Dashboard
Select a market and num of days that affect analysis category and after press analyse button
Mobile Phones
Prognose
1,2,3
Area Chart
Samsung Galaxy Note 4
39,902.00
38,000.00
36,000.00
34,000.00
32,000.00
30,000.00
28,000.00
26,000.00
24,000.00
22,000.00
20,000.00
18,000.00
16,000.00
14,000.00
12,000.00
10,000.00
1
5
10
15
20
25
30
35
37
Donut Chart
View Details
Tasks Panel
Calendar updated
just now
Transactions Panel
Order #
Order Date
Order Time
Amount (USD)


Samsung Galaxy Note 4
39,902.00
38,000.00
36,000.00
34,000.00
32,000.00
30,000.00
28,000.00
26,000.00
24,000.00
22,000.00
20,000.00
18,000.00
16,000.00
14,000.00
12,000.00
10,000.00
1
5
10
15
20
25
30
35
37


Samsung Galaxy Note 4
39,902.00
38,000.00
36,000.00
34,000.00
32,000.00
30,000.00
28,000.00
26,000.00
24,000.00
22,000.00
20,000.00
18,000.00
16,000.00
14,000.00
12,000.00
10,000.00
1
5
10
15
20
25
30
35
37

Висновки

Методи покращення продукту:

1. Розширити модешь на інші цілі
2. Зробити можливість НА.
3. Розробити підтримку багатьох та користувацьких алгоритмів
4. Провести більш глибокий аналіз в варіанті з маркетами.

Загальний результат

В данній роботі виконались всі початкові вимоги з обробки маркетів, була розроблена архітектура що являється універсальною для подібних систем, оцінена кількість вимог до апаратної системи, та розроблено систему з відмовостійкістю.

Дякую за увагу!