

Mining and identifying security threats using Spark SQL, HBase and Solr



Manidipa Mitra

Director, Big Data CoE
ValueLabs



National Security Depository Organization – Statistics

20
years

Data

9
TB

Volume

15
million

Unique Accounts

500 +

Daily Transactions

20
billion

Records in a single table

150 +

Reports

1.5 hours
to **6** hours

Processing Time

Challenges – Cost & Performance



Report 1

Client name and address search report based on unique government identification number

- Used for identifying security
- Any search hits 1.5 TB volume, 50 million records
- 250 simultaneous query searches take 1.5 hours



Report 2

Statement of Holding as on date for a **beneficiary client account**

- Used serve customers' requests as well as action against security threats
- Can hit a table containing up to 2 billion records
- *On a specific instance, the system took 3.5 hours to retrieve 6,594 records*



Report 3

Beneficiary position report

- 60 thousand files to be generated every Friday
- Reports to be generated for 60 million unique records
- Each record contains a few thousand to a few million entries
- Takes 4 hours to generate a report

Query Type

Client and Address Search Report

%joh%do%|%cal%|%23%AF%G

%wen%mar%|%burg%|%56%9%DS

Statement Of Holding Report

XXXXXXXXX A/C number 25thJuly,2016

YYYYYYYYY A/C number 6th Sept 2015

Beneficiary Position Report

Summary weekly report for RTA1

Summary weekly report for RTA2

Environment for PoC

Cluster 1

- 5 Nodes Cluster
- 2 Name Nodes
 - ✓ 8 Cores (2*4), Intel(R)Xeon(R)
 - CPU E5-2670 v3 @ 2.30GHz
 - ✓ 64 GB RAM
- 3 Data Nodes
 - ✓ 16 Cores (2*8), Intel(R) Xeon(R)
 - CPU E5-2670 v3 @ 2.30GHz
 - ✓ 128 GB RAM
 - ✓ 4.8 TB Storage
- Shared by other applications as well

Used for
Report1 and
Report2

Cluster 2

- 5 Nodes Cluster
- 2 Name Nodes
 - ✓ 8 Cores (2*4), Intel(R)Xeon(R)
 - CPU E5-2670 v3 @ 2.30GHz
 - ✓ 32 GB RAM
- 3 Data Nodes
 - ✓ 16 Cores (2*8), Intel(R) Xeon(R)
 - CPU E5-2670 v3 @ 2.30GHz
 - ✓ 48 GB RAM
 - ✓ 1 TB Storage

Used for
Report3

Performance comparison of existing EDW system vs Hadoop

EDW

Hadoop

Report 1

250 simultaneous
queries in 1.5 hours

280 queries : 5 minutes
800 : 12 minutes

Report 2

6,594 records in 3.5
hours

5000 records in ~10 seconds
(10 unique account numbers,
each with 500 records)

Report 3

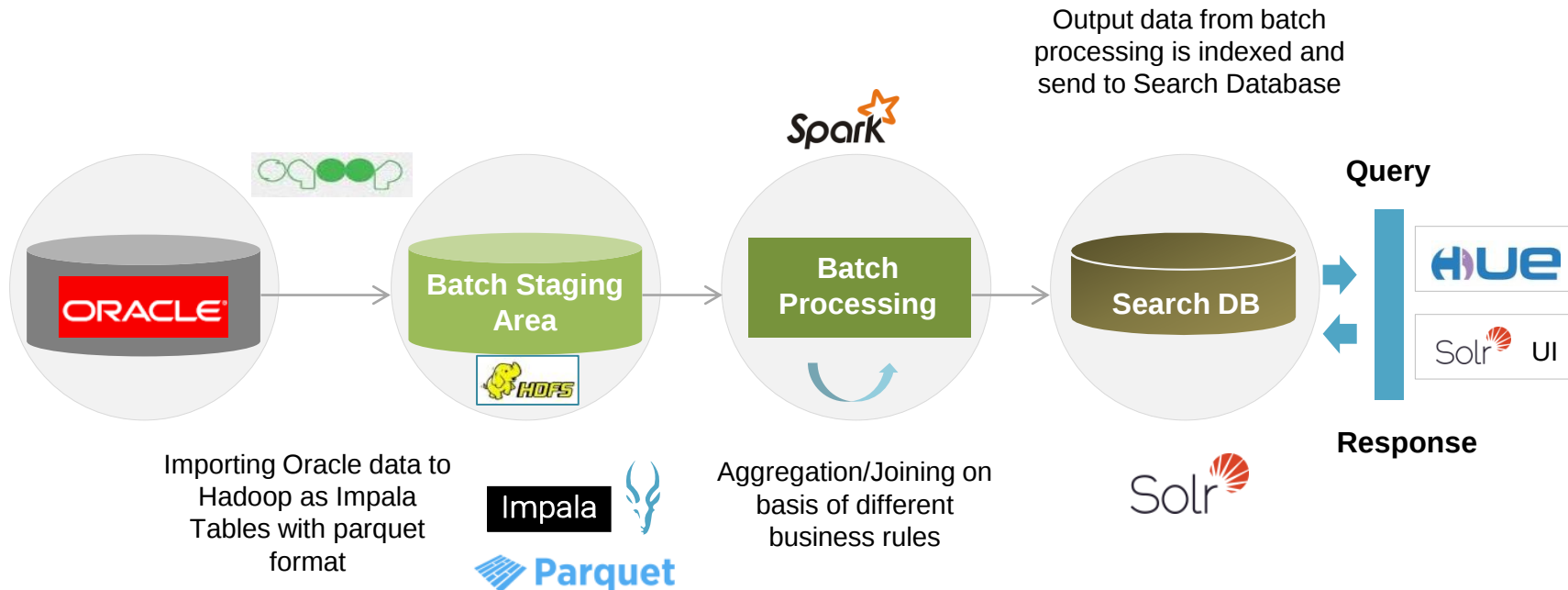
60,000 in 4 hours

10,000 in 10 minutes

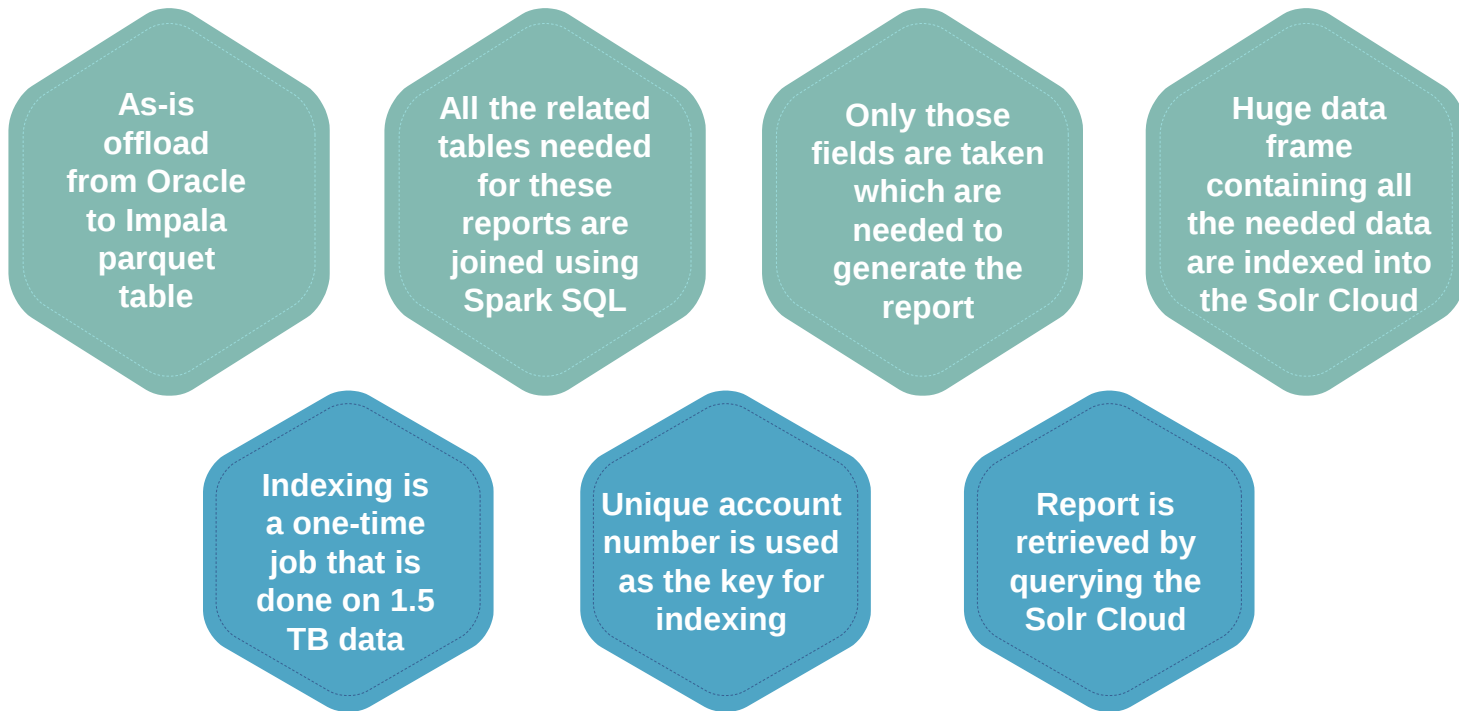
The output (as a report) generated by Hadoop cluster is accurate as EDW

Client name and address search report – Architecture

Achieving 5 mins against 1.5 hours

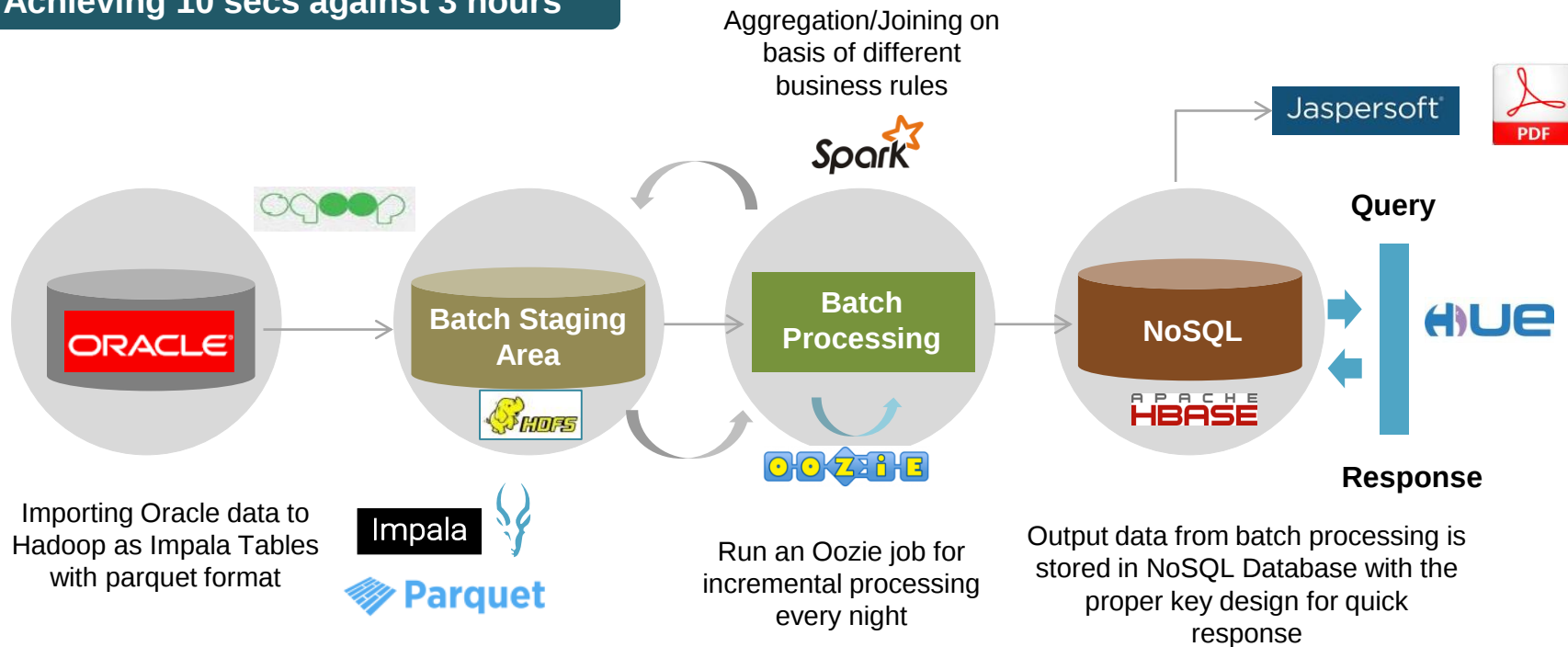


Client name and Address search report(s)



Statement of Holding report generation (in Hadoop way)

Achieving 10 secs against 3 hours



Status of Holding report(s) implementation in Hadoop

Tables are offloaded 'as is' to Hadoop from Oracle in Impala parquet format

All the related tables needed for these reports – Customer Table, Transactions Table, ISIN Tables, Beneficiary, Clearing, etc – are joined using Spark SQL

Created a HBase table 'Agg-Transaction' with a row key as account number | Helps keep the latest SoH value for each unique account number as on date

Created another HBase table 'SoH' with a row key concatenating unique account numbers and the dates with the Status of Holding value and other details required for the report

An Oozie process at 12 a.m. everyday looks into any new transaction(s) on the previous day

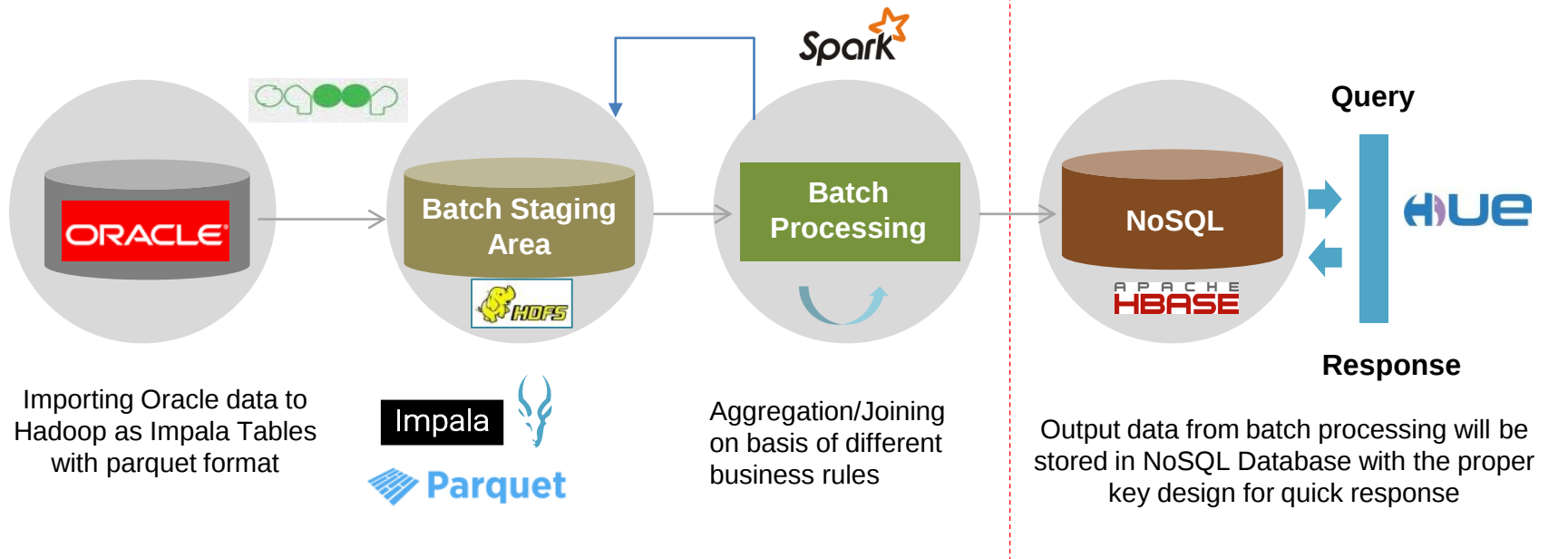
Both tables get updated every night . For new transaction 'Agg-Transaction' got updated, for new share new rows are added on 'Agg-Transaction'. For new shares and transactions new rows are added on SoH

Queries are fired on SoH table to get response in seconds

Beneficiary Position Report Generation (in Hadoop way)

Achieving 1 hr against 4 hours

Output data from batch processing is stored in a HDFS directory as a text file, will be zipped if needed



Beneficiary position report(s) - Achieving in minutes against hours

Tables are offloaded 'as is' to Hadoop from Oracle in Impala parquet format

All the related tables needed for these reports – Customer Table, Transactions Table, ISIN Tables, Beneficiary – are joined using Spark SQL

1 report for each ISIN is generated in a predefined text format

Output is written in standard text format with different fields delimited

Spark cluster is tuned to achieve best performance

Thank You!

There's more to **Value**
than meets the eye.



marketing@valuelabs.com | www.valuelabs.com/marketing