AIOUG

HYDERABAD CHAPTER

25-AUG-2018

HYDERABAD CHAPTER

aioug

ALL INDIA ORACLE USERS GROUP

THE POWER TO CONNECT & CONQUER

APACOUC

WEBINAR TOUR 2018

27-AUG-2018

APACOUC

APAC ORACLE USERS GROUPS COMMUNITY

SURESH GANDHI

Oracle Cloud:
Building Data lake

About Me

- Oracle DBA since Version 7
- 17 Years experience in Database Technology
- Happy DBA until NOSQL arrived
- Then NOSQL Database Knowledge
- Pleasant DBA until Cloud arrived, and then learnt Cloud Technologies
- Oracle/Redhat/AWS/ITIL Certified
- Technology Director for Independent Australia Oracle User Group
- Blogger http://db.geeksinsight.com

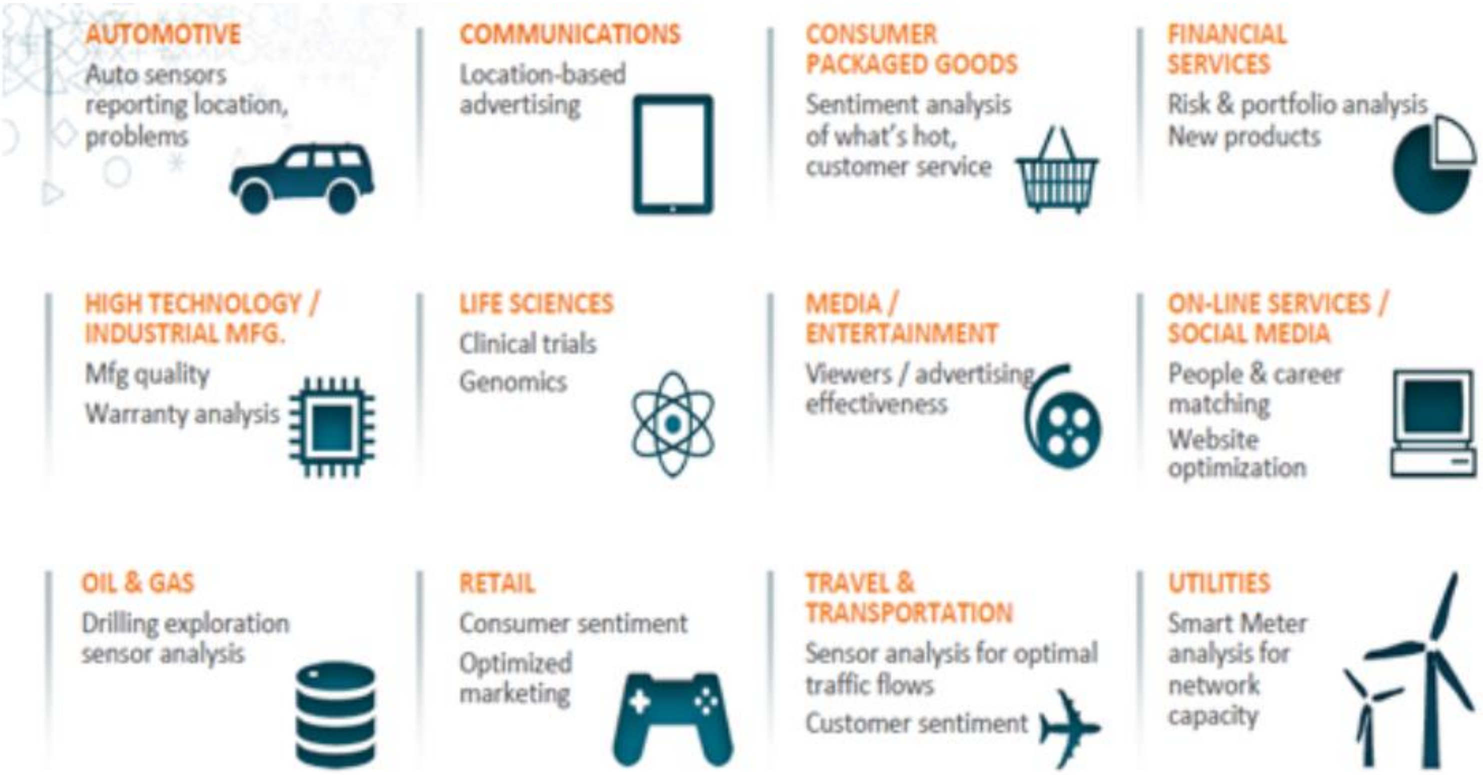
Today’s agenda

- What is a Datalake?
- Why Datalake?
- Data Storage Patterns
- ETL vs ELT
- DatawareHouse vs. Datalake
- Is Datalake means only Big Data Solution?
- Considerations for Building a Datalake
- Data Lake Architecture & Services
- Strategies : Building Blocks of a Data Lake
- Datalake in Oracle Cloud

Why a datalake

The BIG DATA world

- We are now in the BIGDATA era and data is everywhere. The sources and formats of Data is different than earlier days.
- We need a platform for managing & consuming Data
- We need complex analysis to do than before
- Business Models often change these days and its relative data models too



AUTOMOTIVE
Auto sensors reporting location, problems

COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, customer service

FINANCIAL SERVICES
Risk & portfolio analysis
New products

HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg quality
Warranty analysis

LIFE SCIENCES
Clinical trials
Genomics

MEDIA / ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching
Website optimization

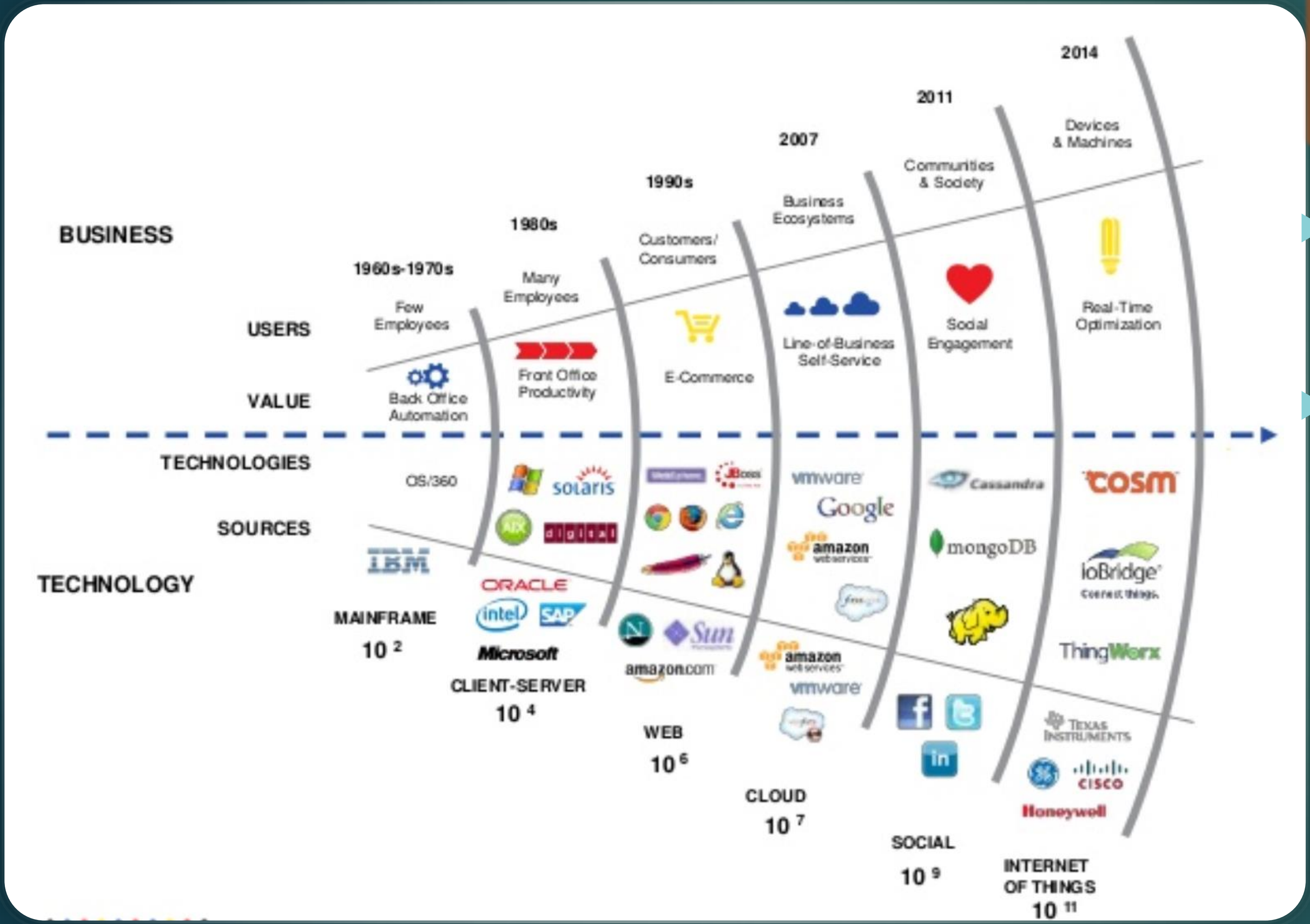
OIL & GAS
Drilling exploration sensor analysis

RETAIL
Consumer sentiment
Optimized marketing

TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows
Customer sentiment

UTILITIES
Smart Meter analysis for network capacity

Data challenge



- Changes in the business patterns introduces changes in technologies
- This introduced changes in data sources and data patterns and data volumes

Image Source: informatica

Why Not datawarehouse?

Datawarehouse

Works on specific data and mine through and provide subsets of data

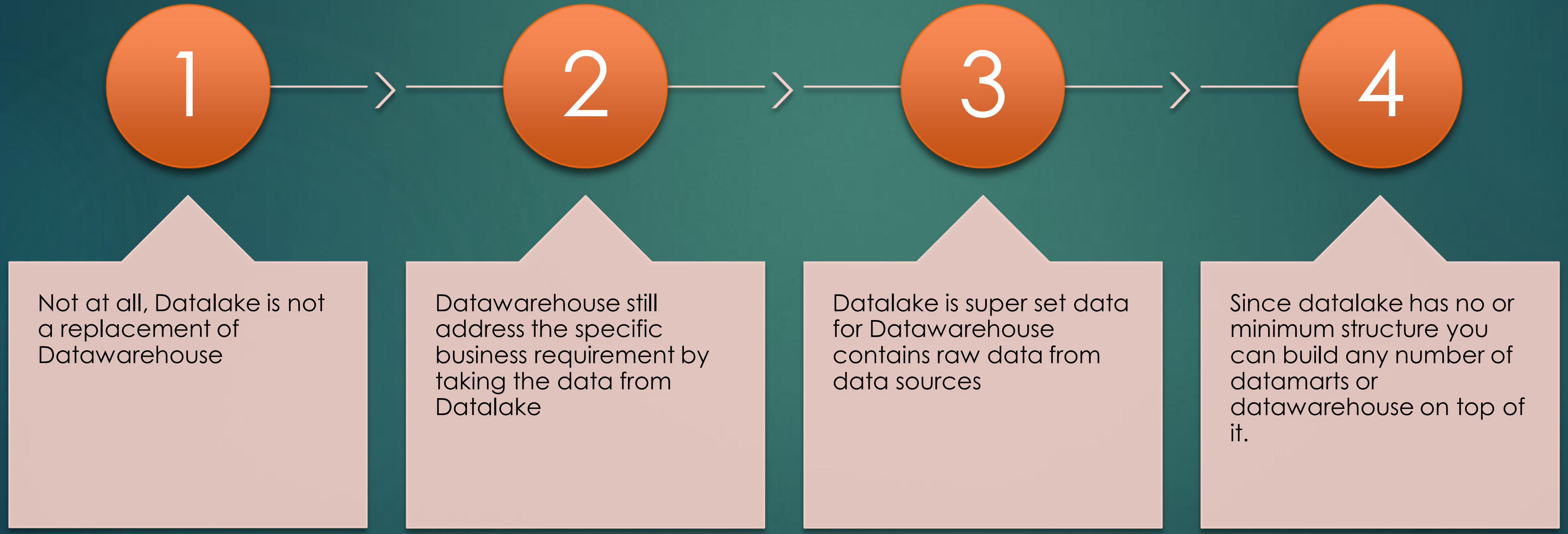
DataLake

Works on variety of data and able to run complex predictive analysis on data

Datawarehouse Vs. DataLake

	Datawarehouse	DataLake
Storage	Storage is usually on expensive Tier 1 which helps ensure performance to query data, availability and backup. Generally we are talking about a database	Storage is usually a Hadoop cluster which is designed for storing vast amounts of data for big data solutions at very low cost.
Data Access	Schema on write Data structures in the data warehouse need to be modelled before data can be loaded. Access via standard SQL and BI tools.	Schema on read No predefined data structures are needed. Structures will be created when and if needed for processing. Access via SQL like languages (like Hive and Pig) or batch processing (like MapReduce).
Security	Well Defined Security for database and file systems are well established and mature. It can allow individuals or groups of users to access all or just a portion of data.	Partially or not defined The security mechanisms for these solutions are still maturing. A trade-off between security restrictions on data and making all data available for data discovery needs to be found.
Data Lineage	Defined Data can be traced from source to the data warehouse via ETL and mapping processes.	Not defined Lineage is not defined since data is transformed and processed in many different ways as needed.
Data Governance	Defined and well established All data loaded has an owner, a load frequency, clear lineage, a defined business purpose, security and compliance policies, relevant masking and encryption.	Partially defined and still evolving Governance is not a question specifically relating to data lakes but to big data in general. Policies still need to be defined that clarify the minimum set of governance rules that should apply to this type of architecture.
Data Quality	Enforced Data loaded is cleansed, formatted and loaded or altered according to consistent rules and processes	Evolving Processes like ETL, cleansing, matching, de-duping, merging, parsing and standardising are still evolving in the big data / Hadoop environment.
Data and Types	Structured, cleansed, processed and transformed Database data Flat files Xml files JSON	Raw untransformed, semi- structured and unstructured Examples: Clickstream data, Logs, Emails, Social media, Audio and video files and Geolocation coordinates.
Flexibility	Rigid All data structures and loading processes are predefined. Any change to a data object or process needs careful consideration for impact of existing solution and requires a considerable amount of time.	Very flexible Since there are no specific data structures in place it is easy to use the data lake whichever way the data scientists and analysts see fit. Models can be easily modified, trained and retrained in many different ways allowing for different outcomes to be explored.
Usage	Static and ad hoc reporting Only a subset of attributes is loaded and considered, based on the data and information necessary to satisfy the business needs, this means that only predetermined questions can be asked. Aggregation may cause a loss of information.	Analytics All existing data is loaded .This means that any question can be asked, for example: Improving web advertising via clickstream data; or Improving understanding of customer behaviour via social media data.

What about our datawarehouse, do we replace ?



1

Not at all, Datalake is not a replacement of Datawarehouse

2

Datawarehouse still address the specific business requirement by taking the data from Datalake

3

Datalake is super set data for Datawarehouse contains raw data from data sources

4

Since datalake has no or minimum structure you can build any number of datamarts or datawarehouse on top of it.

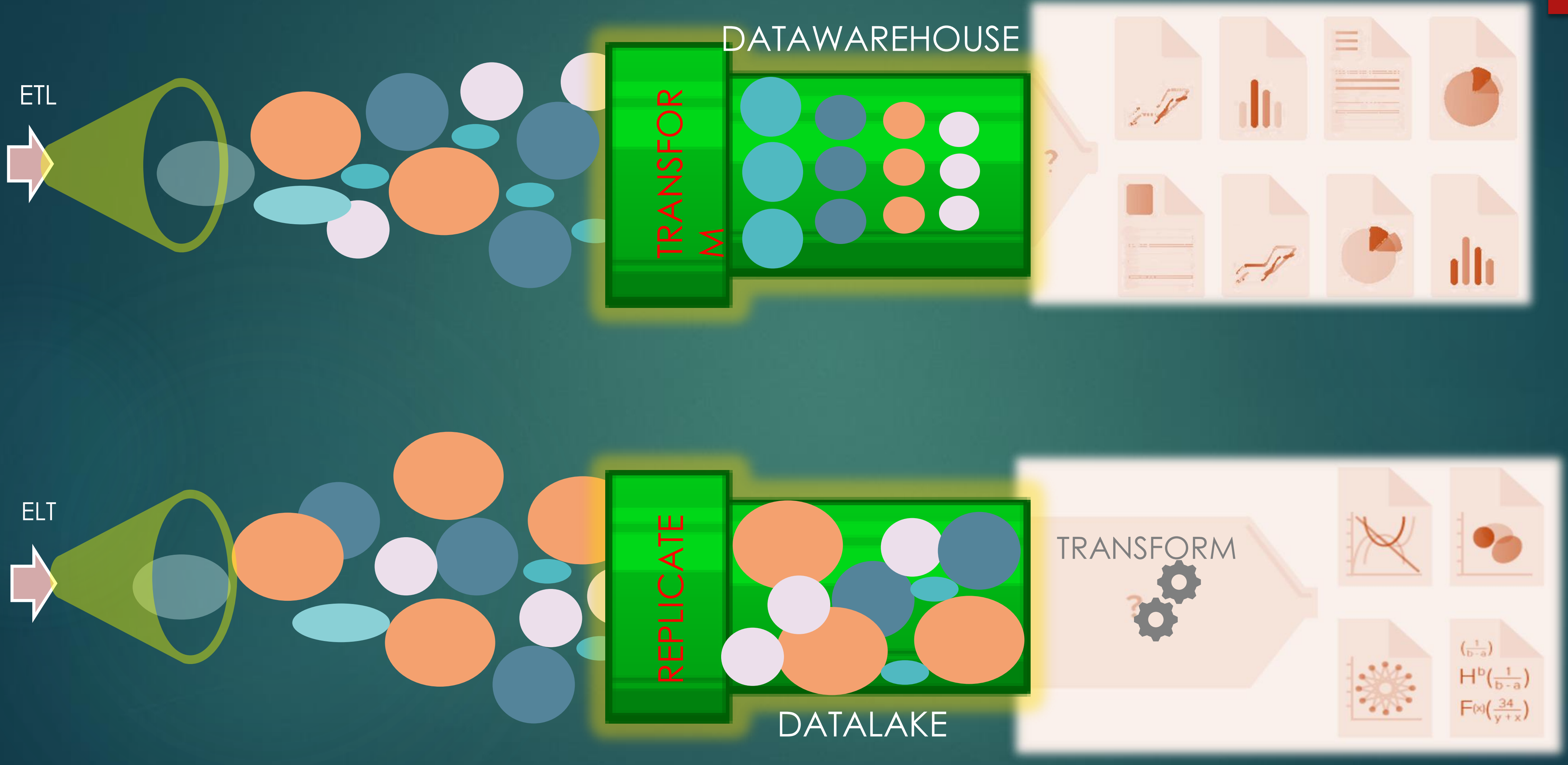
ELT vs. ETL

DataLake uses

ELT over ETL

- ETL is normally a continuous, ongoing process with a well-defined workflow. ETL first extracts data from homogeneous or heterogeneous data sources. Then, data is cleansed, enriched, transformed, and stored either back in the lake or in a data warehouse.
- ELT (Extract, Load, Transform) is a variant of ETL wherein the extracted data is first loaded into the target system. Transformations are performed after the data is loaded into the data warehouse. ELT typically works well when the target system is powerful enough to handle transformations. Analytical databases like Amazon Redshift and Google BigQuery are often used in ELT pipelines because they are highly efficient in performing transformations.

DATALAKE – WHAT METHOD TO CHOOSE?
ETL VS ELT



ETL

TRANSFOR
M

DATAWAREHOUSE

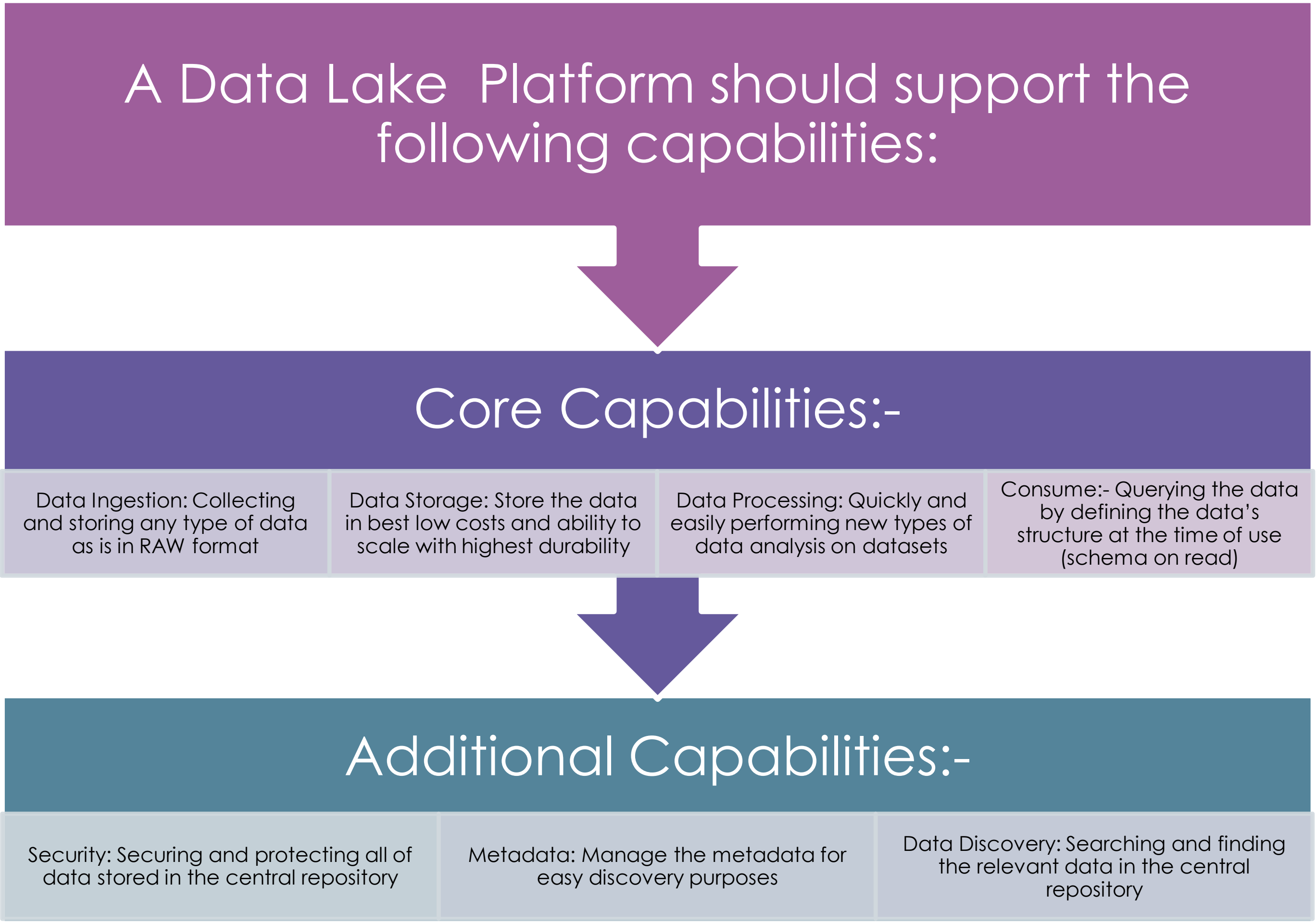
ELT

REPLICATE

DATALAKE

TRANSFORM

Must have Datalake capabilities



A Data Lake Platform should support the following capabilities:

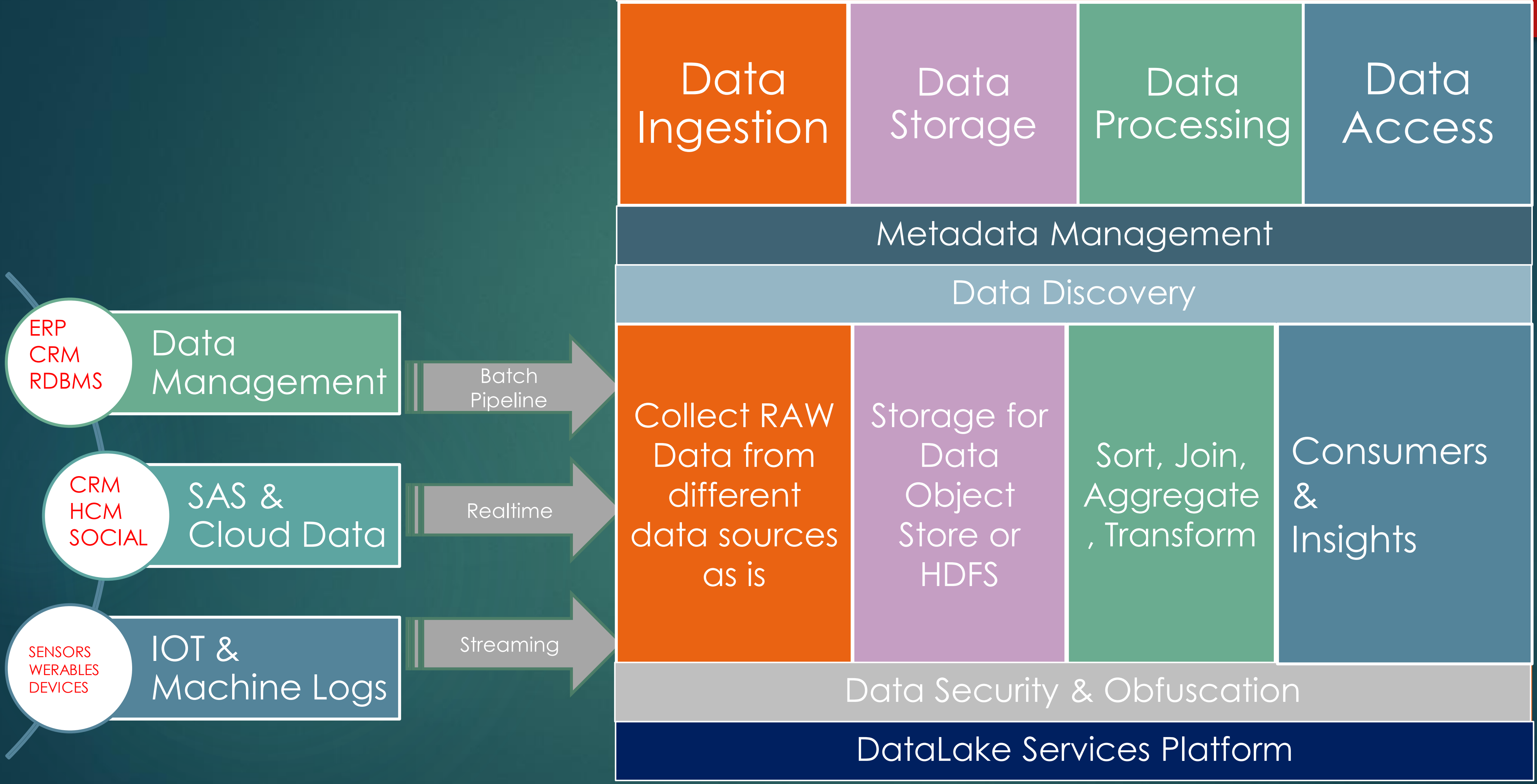
Core Capabilities:-

- Data Ingestion: Collecting and storing any type of data as is in RAW format
- Data Storage: Store the data in best low costs and ability to scale with highest durability
- Data Processing: Quickly and easily performing new types of data analysis on datasets
- Consume:- Querying the data by defining the data's structure at the time of use (schema on read)

Additional Capabilities:-

- Security: Securing and protecting all of data stored in the central repository
- Metadata: Manage the metadata for easy discovery purposes
- Data Discovery: Searching and finding the relevant data in the central repository

DATALAKE ARCHITECTURE



Data Ingestion

Data Storage

Data Processing

Data Access

Metadata Management

Data Discovery

Collect RAW Data from different data sources as is

Storage for Data Object Store or HDFS

Sort, Join, Aggregate , Transform

Consumers & Insights

Data Security & Obfuscation

DataLake Services Platform

ERP
CRM
RDBMS

Data Management

Batch Pipeline

CRM
HCM
SOCIAL

SAS & Cloud Data

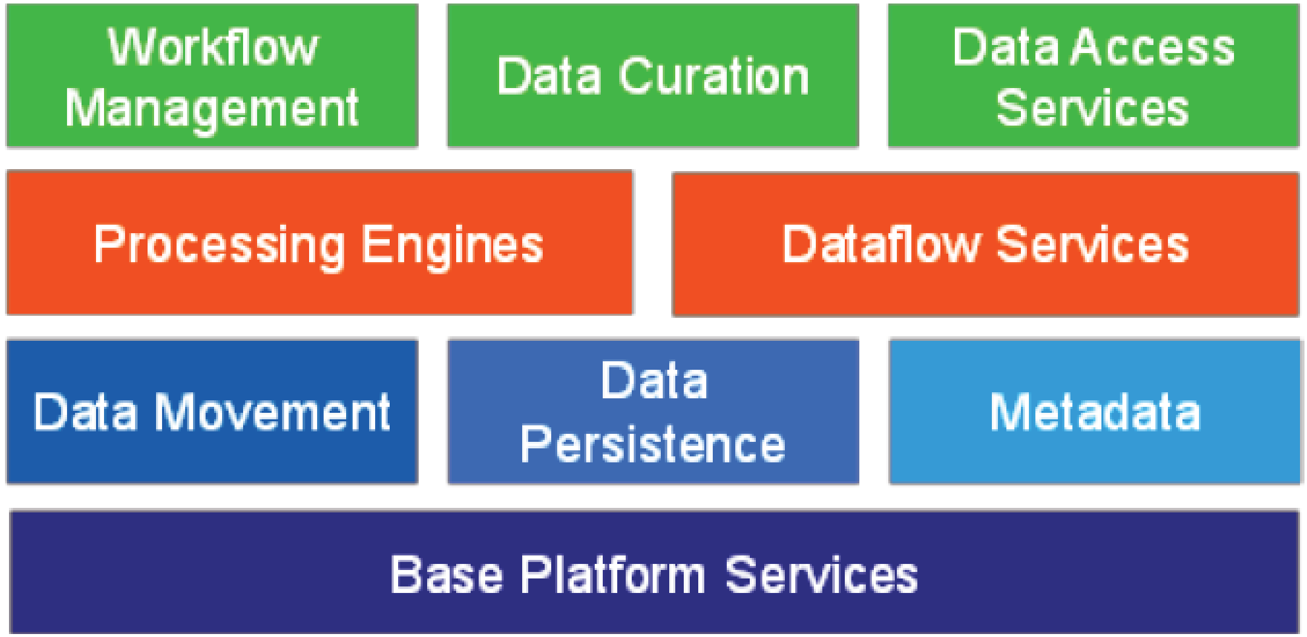
Realtime

SENSORS
WERABLES
DEVICES

IOT & Machine Logs

Streaming

Core services



Workflow Management

Data Curation

Data Access Services

Processing Engines

Dataflow Services

Data Movement

Data Persistence

Metadata

Base Platform Services

Technology architecture

Right Platform + Right Tools + Right Interface

Generally Datalake’s built on Hadoop Eco System

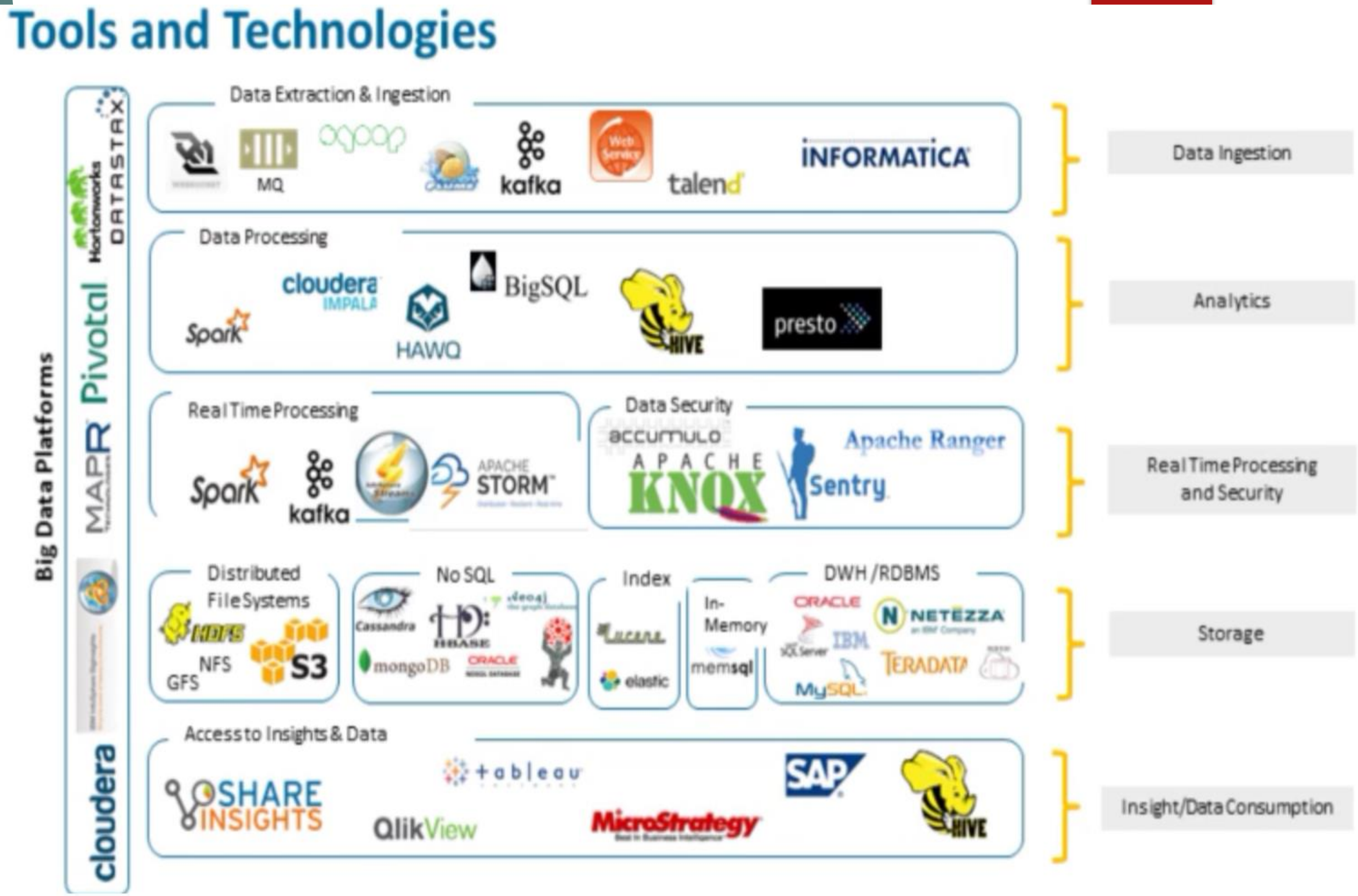
But

I will leave this slide here, will not discuss what is right since in your case depends on your organisational goals the Platform , Tools, Interface may completely different than what rest of world use

So I urge , you must evaluate multiple platforms, multiple tools & interfaces and conduct POC before jump into datalake platform

Data platform tools & technologies

- Generally around Big Data Technologies
- Ingestion, Storage, Processing& Consumption are the building blocks of any datalake
- Governance & Managing Metadata is complementary and should be part of data services



Datalake in Oracle Cloud

Data Engineeer's
Vs
Data Scientists

- "Data engineers" can code, run clusters, and so on, in support of what's always been called "data science". Their knowledge of the math of machine learning/predictive modeling and so on may, however, be limited.

"Data scientists" can write and run scripts on single nodes; anything more on the engineering side might strain them. But they have no-apologies skills in the areas of modeling/machine learning."

"Like data scientists, data engineers write code. They're highly analytical, and are interested in data visualization.

Unlike data scientists—and inspired by our more mature parent, software engineering—data engineers build tools, infrastructure, frameworks, and services. In fact, it's arguable that data engineering is much closer to software engineering than it is to a data science

Source: Mark Rittman Blog

DO YOU NEED HADOOP ECO SYSTEM FOR YOUR DATALAKE?

Well it depends

- Predictive analysis, Machine Learning, Complex analytics requires lot of processing capacity
- When you have variety of data, volume of data Hadoop works well
- But when you really , have more SQL based queries in your organisation and working with smaller sets of data then Hadoop is not right choice.
- You need to decide on which model you need.
- If you organisation has a shifting paradigm in data services model and building datalake you may need to choose to build hybrid model
  - Ingest raw data as is
  - Store the data in flat filesystem or object store (s3, hdfs etc.) and also convert as much as data possible to a database as is.
  - Process
    - For Predictive analysis model you can use Hadoop type processing
    - For general adhoc and support queries or reporting use database

Object store Is a new datalake

A data lake is a key element of any big data strategy and conventional wisdom has it that Hadoop/HDFS is the core of your lake. But conventional wisdom changes with new information and in this case that new information is all about Object Storage. There are many ways to persist data in cloud platforms today such as Object, Block, File, SMB, DB, Queue, Archive, etc. As an overview, here are Oracle's, AWS' and Azure's primary storage solutions.

Object Based Distributed Storage:

- Key/Content driven interface
- Oracle Object Store
- AWS S3
- Azure Blob Storage

File Based Distributed Storage:

- Nested file/folders interface
- Oracle BDCS-CE Storage (HDFS)
- AWS EMR HDFS/EMRFS
- Azure Data Lake Store (HDFS)

Block Based Storage:

- Raw disk like 1s and 0s interface
- Oracle Cloud Block Volume Storage
- AWS Elastic Block Storage (EBS)
- Azure Disk Storage

Object Storage decouple the storage from Compute Engines and its scalable and durable with lower cost

Oracle Cloud

DATALAKE SOLUTIONS & PATTERNS

Data Science Lab Solution Pattern

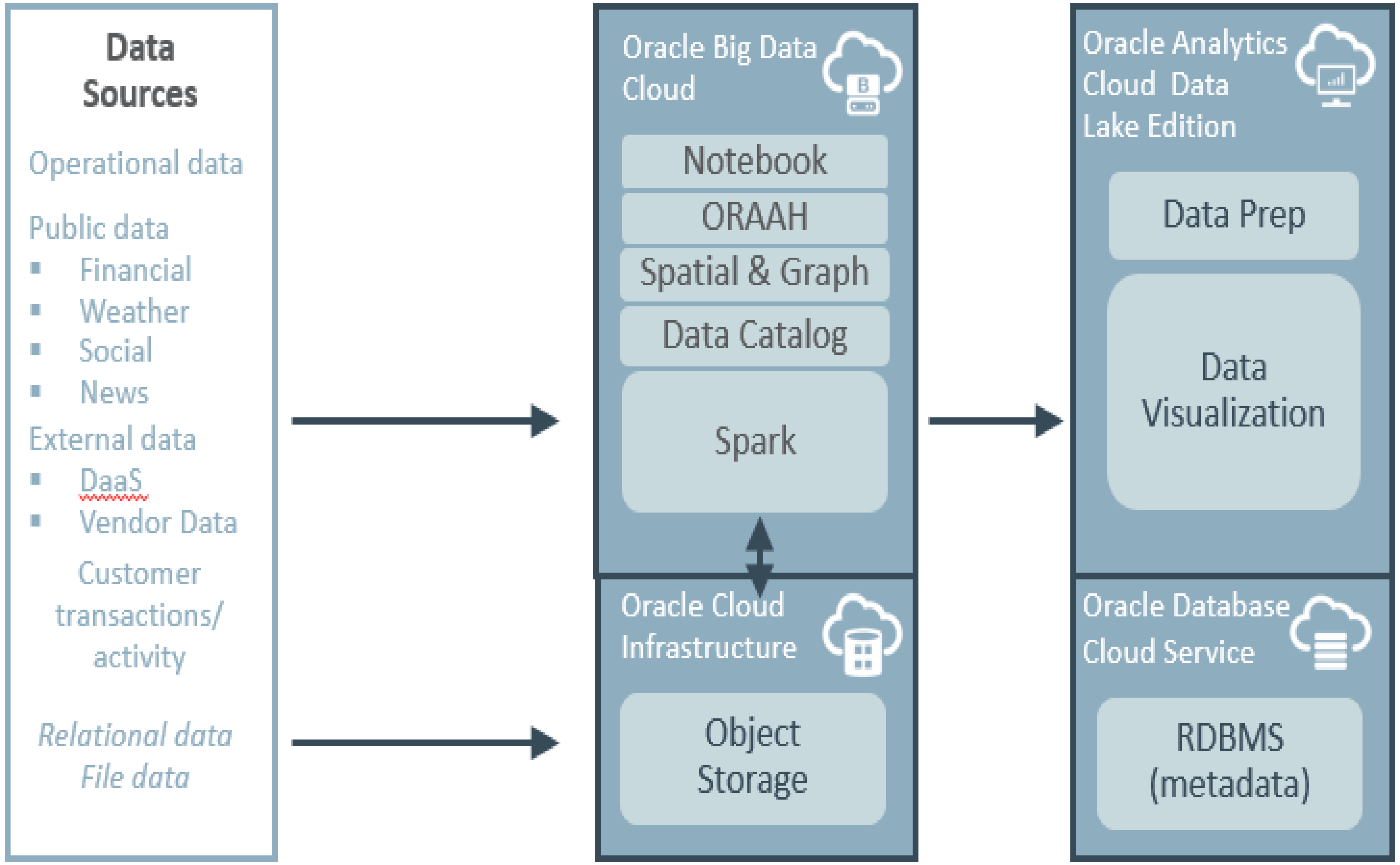


Image Source: Oracle Documentations

ETL Offload for Data Warehouse Solution Pattern

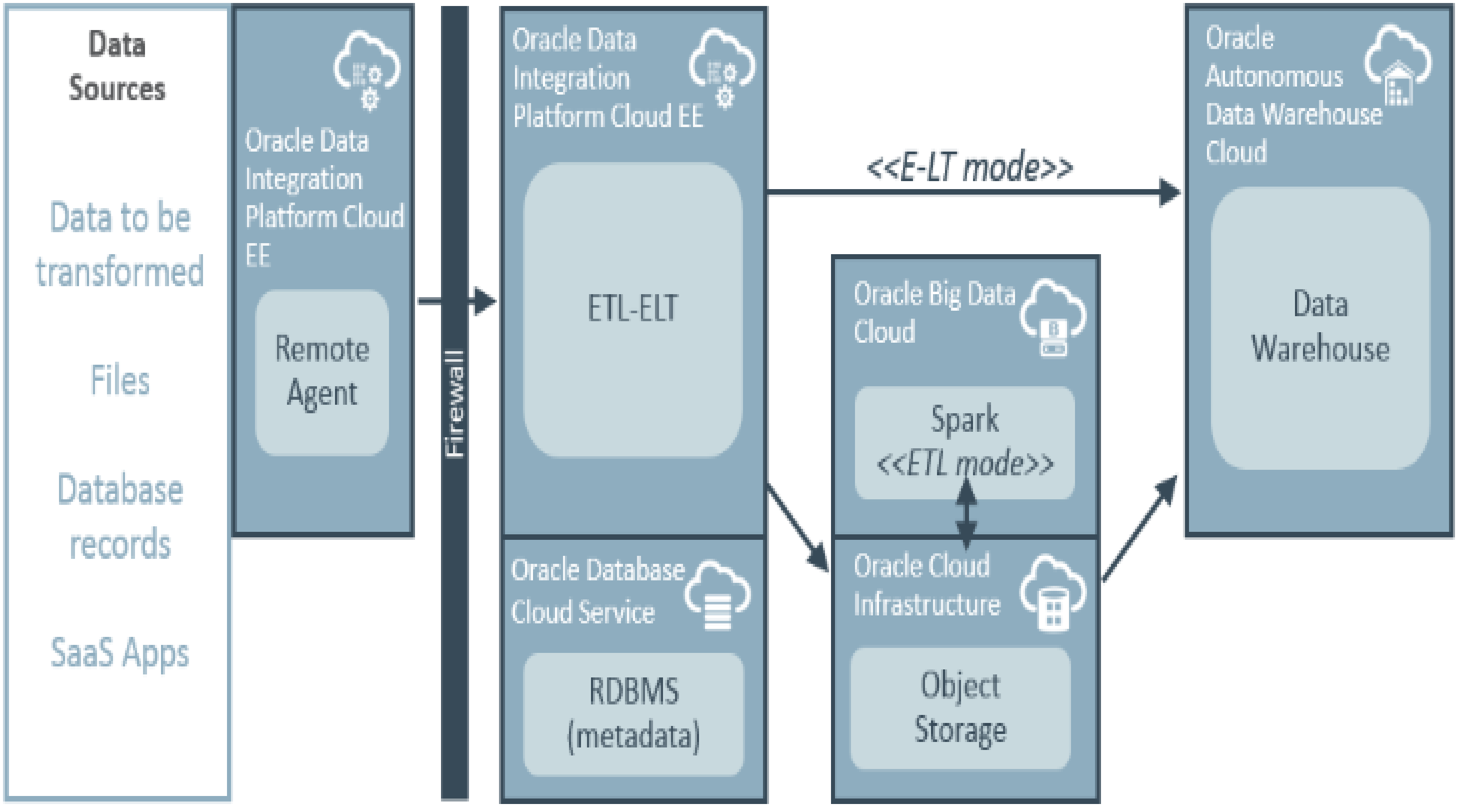


Image Source: Oracle Documentations

Big Data Advanced Analytics Solution Pattern

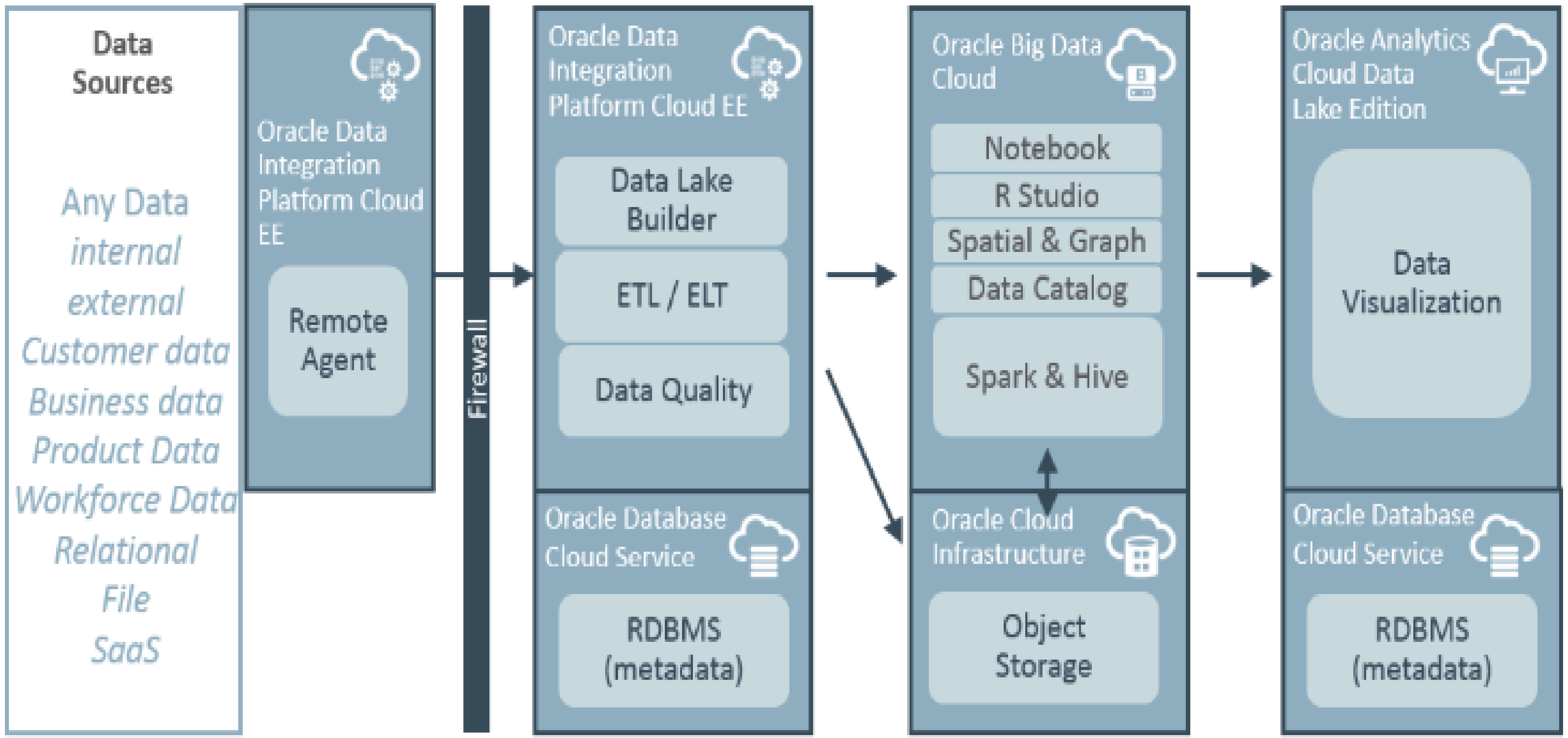


Image Source: Oracle Documentations

Stream Analytics Solution Pattern

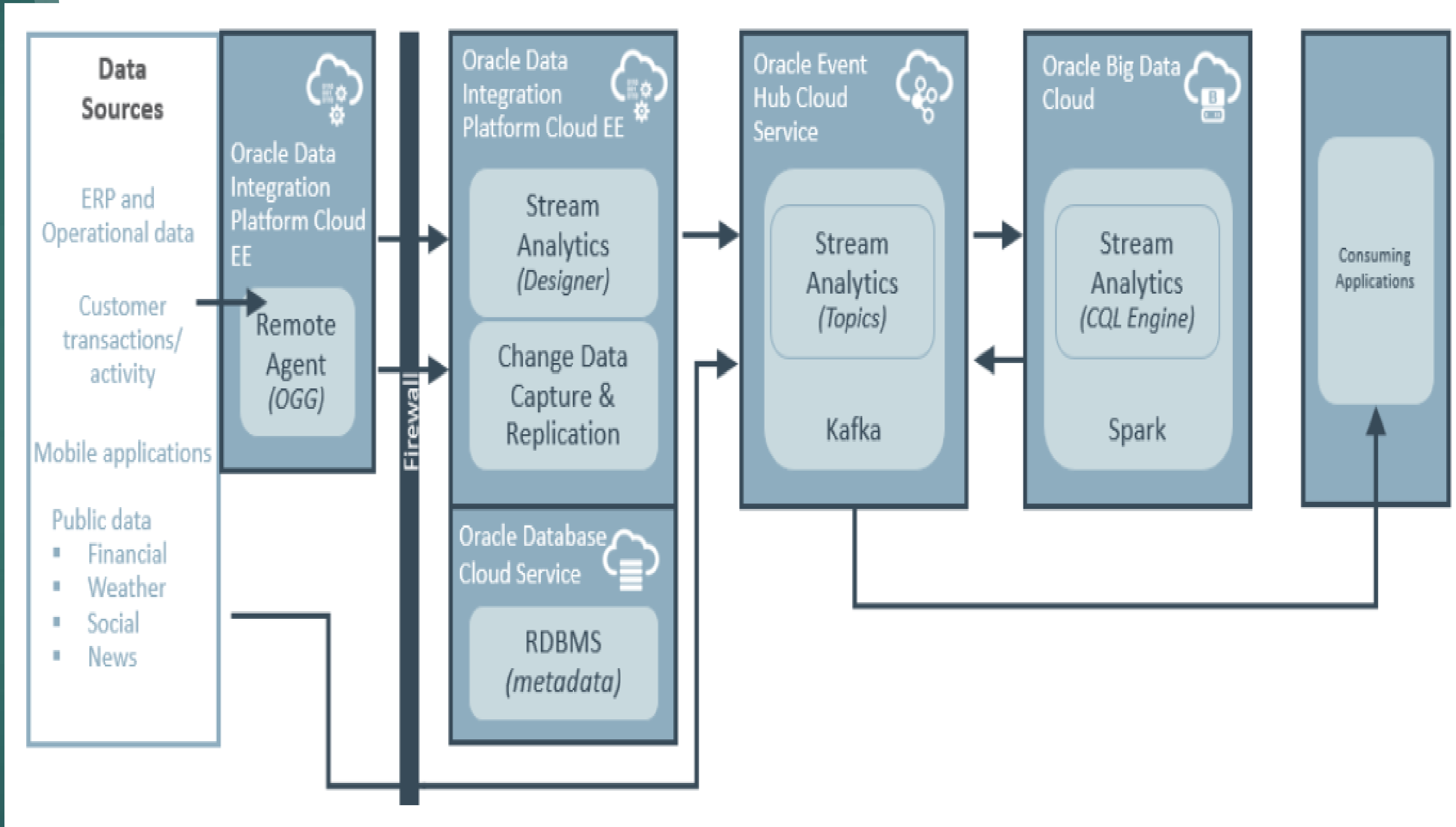
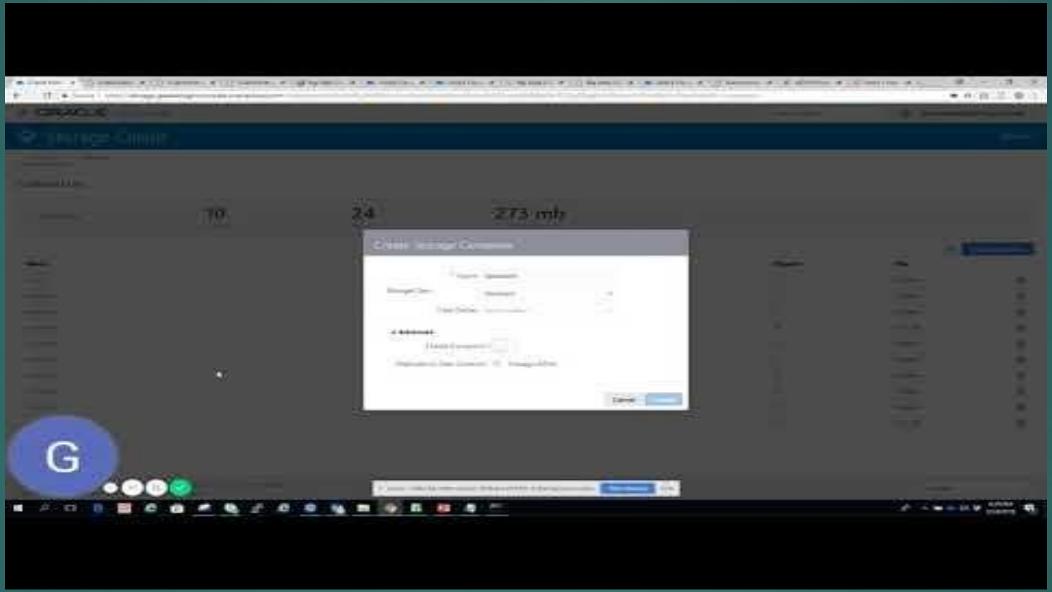

Data Sources
ERP and Operational data
Customer transactions/ activity
Mobile applications
Public data
▪ Financial
▪ Weather
▪ Social
▪ News
Oracle Data Integration Platform Cloud EE
Remote Agent (OGG)
Firewall
Oracle Data Integration Platform Cloud EE
Stream Analytics (Designer)
Change Data Capture & Replication
Oracle Database Cloud Service
RDBMS (metadata)
Oracle Event Hub Cloud Service
Stream Analytics (Topics)
Kafka
Oracle Big Data Cloud
Stream Analytics (CQL Engine)
Spark
Consuming Applications

Image Source: Oracle Documentations

Cont..

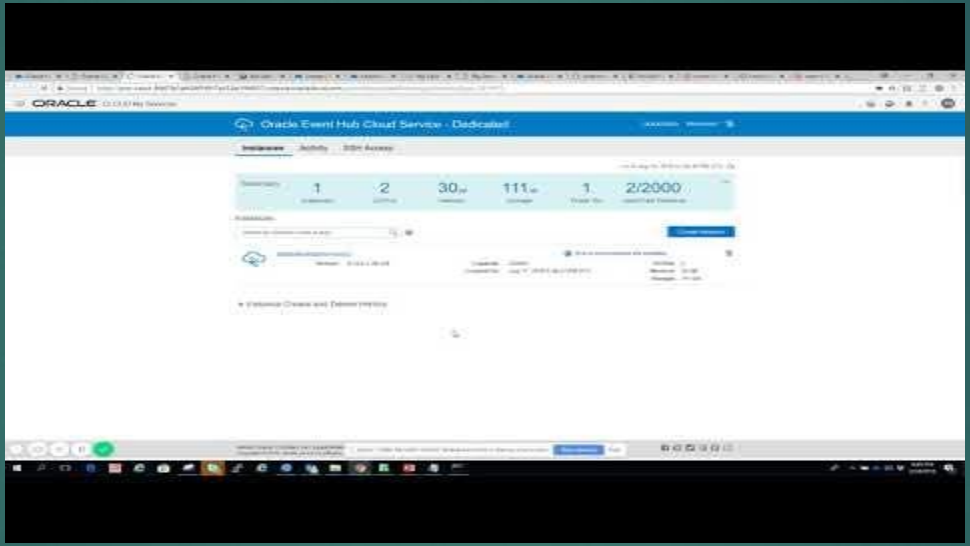
- The development environment I put together for this scenario used the following Oracle Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) components:
- **Oracle Compute Classic** and **Storage Classic** Services
- **Oracle Database Cloud Service**, with the 11g database option
- **Oracle Event Hub Cloud Service Dedicated**, with Kafka Connect and REST Proxy nodes
- **Oracle Big Data Cloud**, single node with Hive, Spark 2.1, Tez, HDFS, Zookeeper, Zeppelin, Pig and Ambari
- **Oracle Autonomous Dataware House (optional)**
- **Oracle Analytics Cloud Data Lake Edition** with Self Service Data Preparation, Visualisation and Smart Discovery (aka Oracle DV)

Introduction Object Store



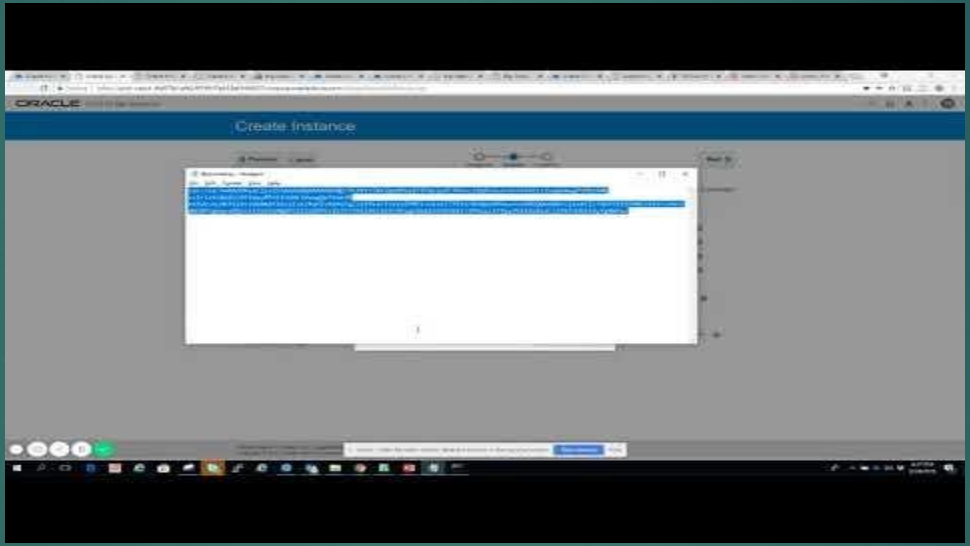
https://youtu.be/L6Tz66J77G8

INTRODUCTION: ORACLE EVENT HUB SERVICE



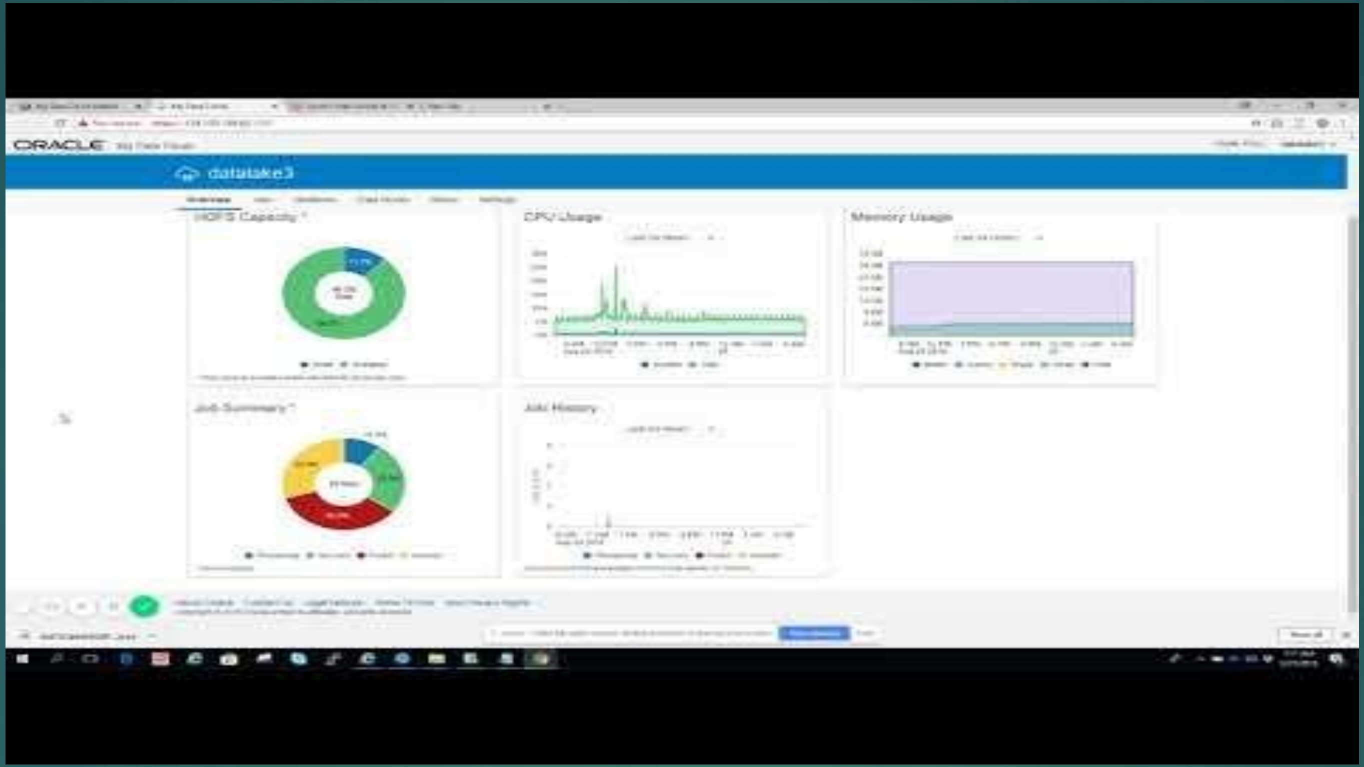
https://youtu.be/YRQWxXZNLzY

INTRODUCTION : ORACLE DATABASE CLOUD



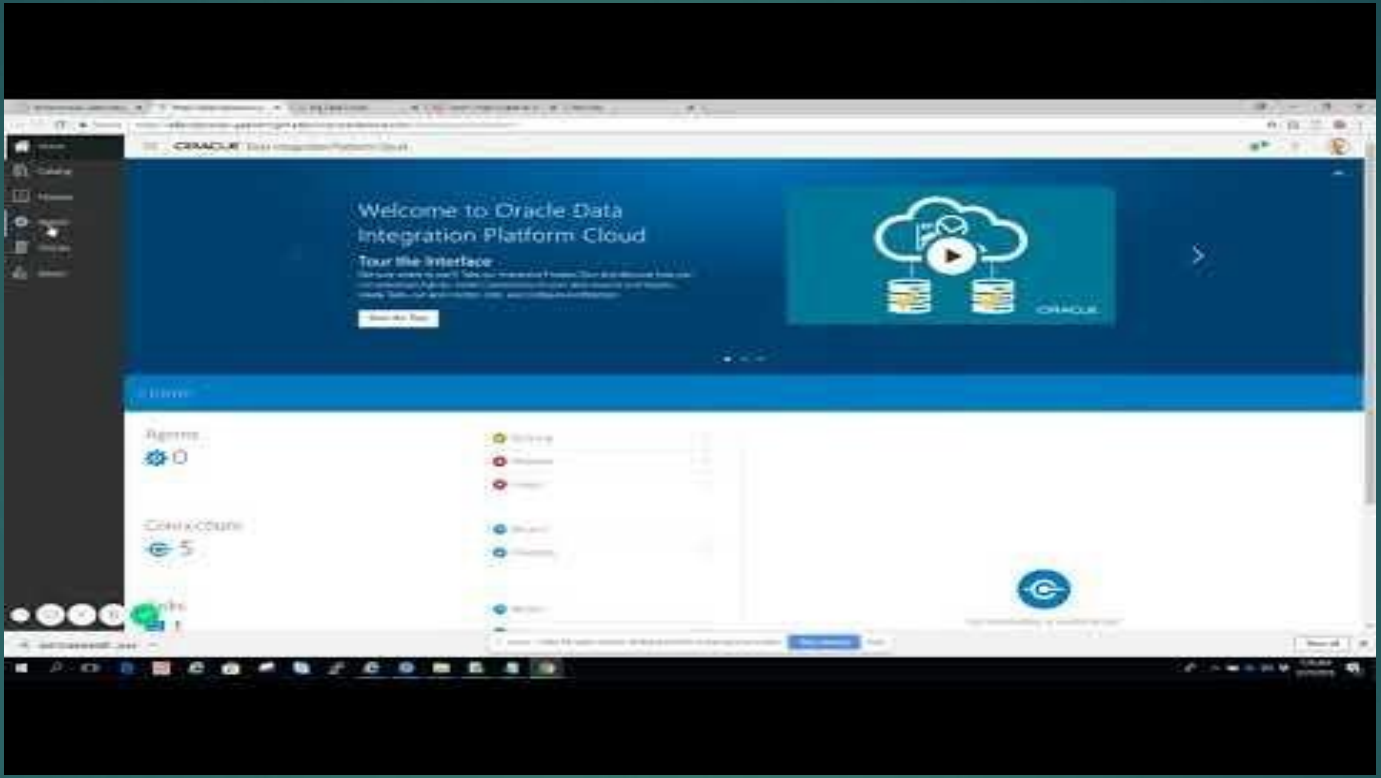
https://youtu.be/jTSdra5uEK4

INTRODUCTION : ORACLE BIG DATA CLOUD



https://youtu.be/zRcJFFLjqAc

INTRODUCTION : ORACLE DATA INTEGRATION PLATFORM CLOUD

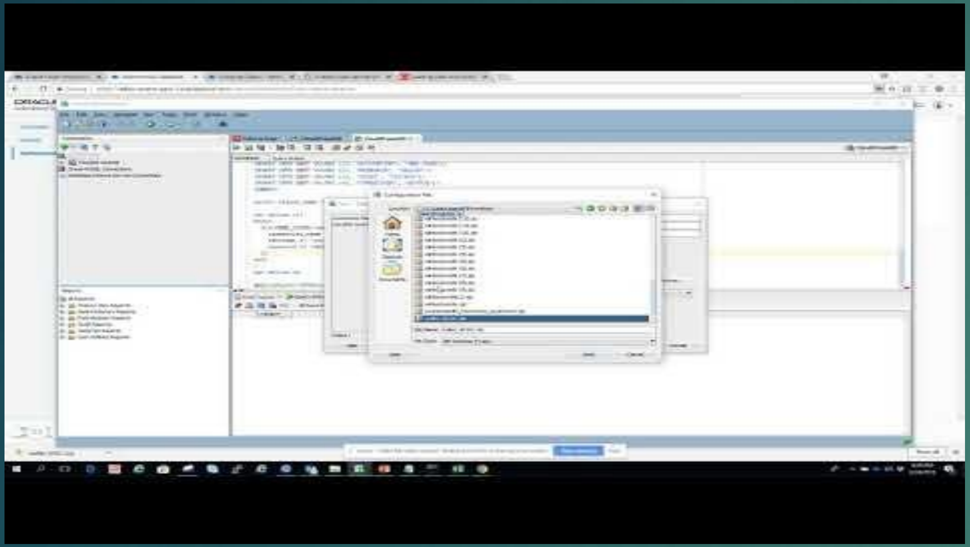


Welcome to Oracle Data Integration Platform Cloud

Tour the Interface

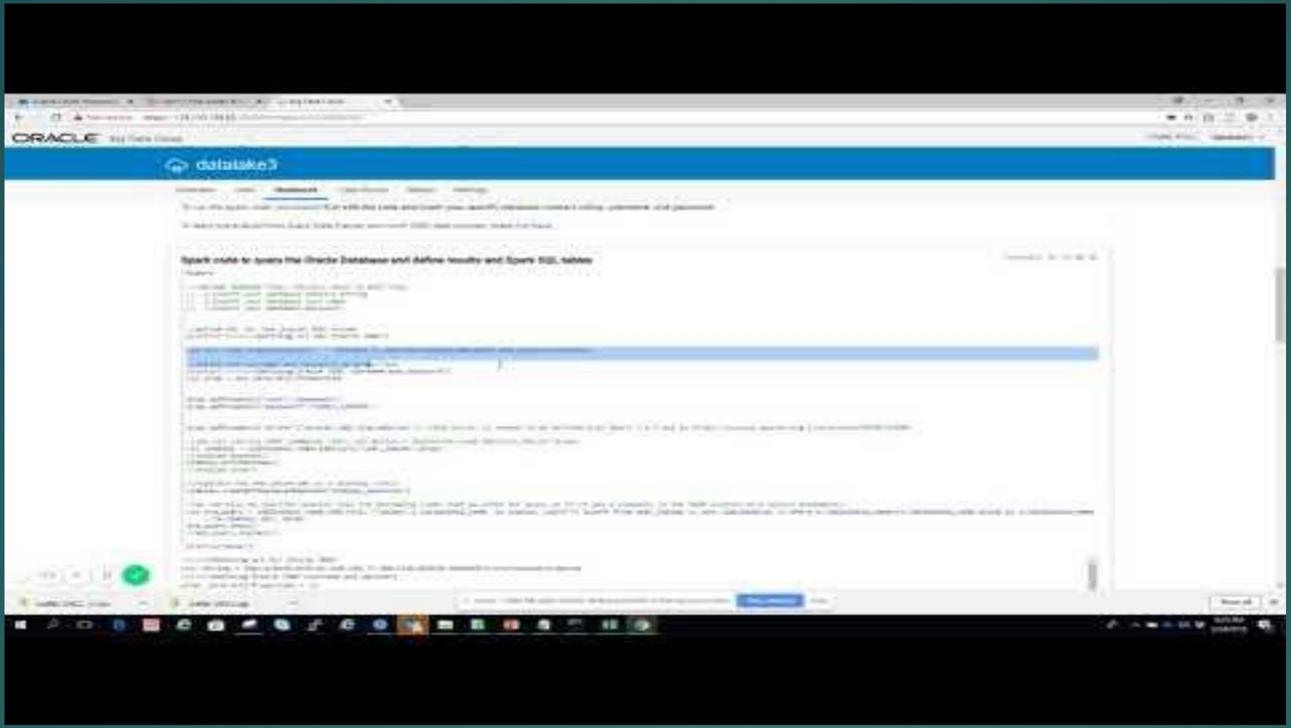
https://youtu.be/oAydpTAu_k4

INTRODUCTION : ORACLE AUTONOMOUS DATAWARE HOUSE



INTRODUCTION : AUTONOMOUS DW – COPY DATA FROM DATALAKE OR OBJECT

INTRODUCTION : ORACLE DATA DISCOVERY

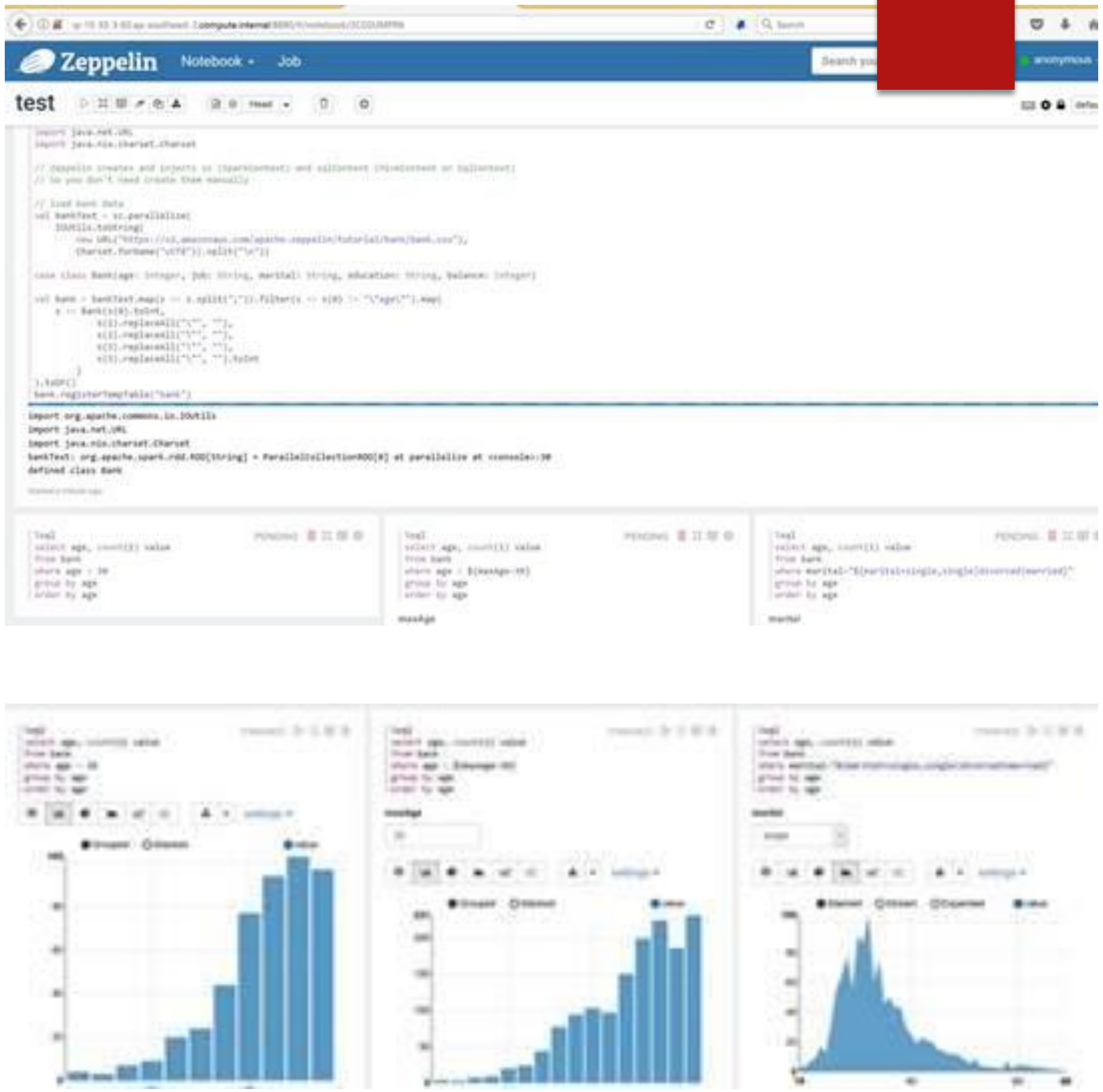


Spark code to query the Oracle Database and define results and Spark SQL tables

https://youtu.be/WV7s4NC7nWc

Data discovery : notebook tools

- Developer can use tools like Apache Zeppelin for development
- Spark, Python, RDD, R scripts
- Pyspark to read flat files and build data frames
- Harmonise the data using r
- Present the end result as flat file again or send to datahub (datawarehouse like redshift)
- Apache Drill
- AWS Athena



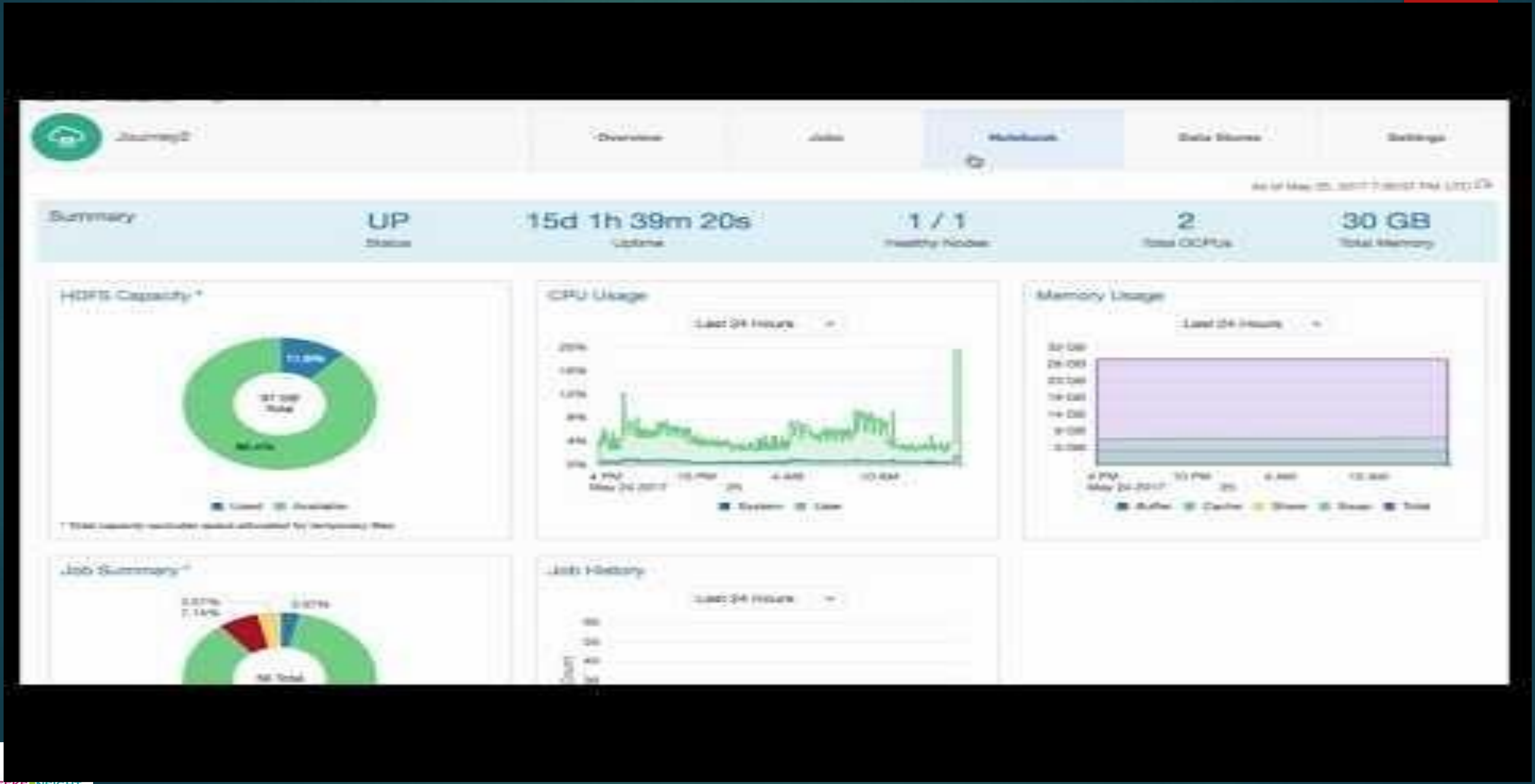
```
import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset
bankText: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[0] at parallelize at <console>:30
defined class Bank
```

```
%sql
select age, count(1) value
from bank
where age < 30
group by age
order by age
```

```
%sql
select age, count(1) value
from bank
where age < ${maxAge=30}
group by age
order by age
```

```
%sql
select age, count(1) value
from bank
where marital="${marital=single,single|divorced|married}"
group by age
order by age
```

Analysing Data in Oracle Data lake



Summary

UP
Status

15d 1h 39m 20s
Uptime

1 / 1
Healthy Nodes

2
Total OCPUs

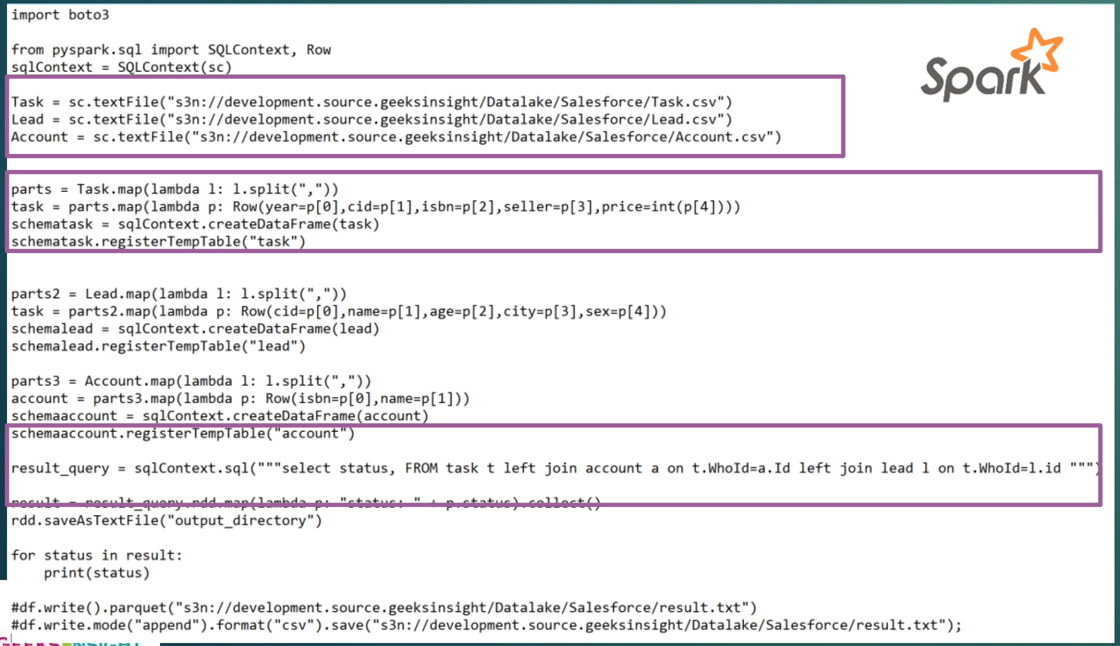
30 GB
Total Memory

Sample video: Analyze data in aws datalake

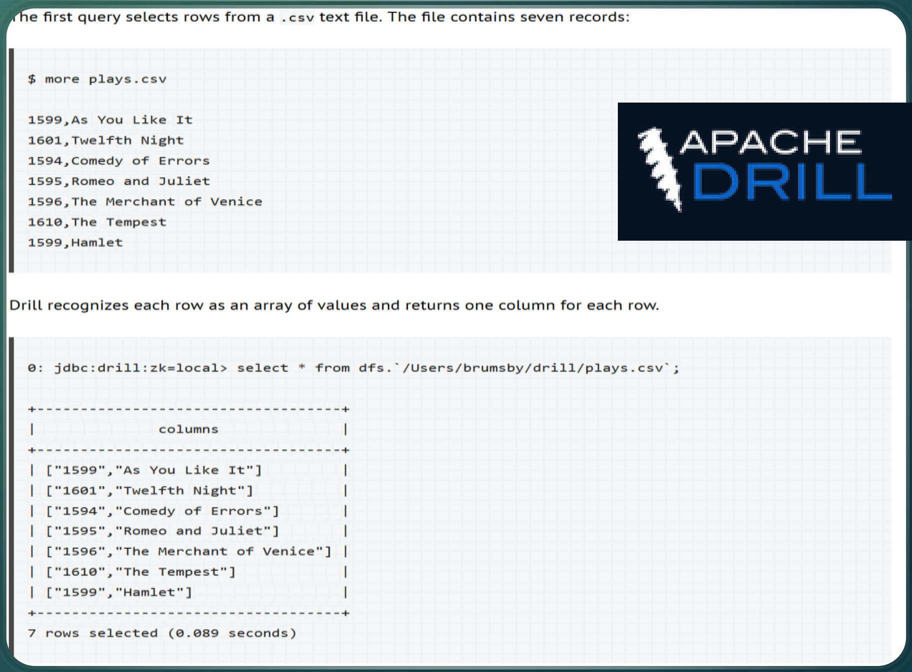
Most commonly asked question

- Can I run a sql query on Object Store, DataLake
- How do I run a query on Flat File
- How do I join two files

1. Apache Drill
2. Pyspark
   1. SQLContext
   2. DataFrames
3. Hive External Tables



```
import boto3

from pyspark.sql import SQLContext, Row
sqlContext = SQLContext(sc)

Task = sc.textFile("s3n://development.source.geeksinsight/Datalake/Salesforce/Task.csv")
Lead = sc.textFile("s3n://development.source.geeksinsight/Datalake/Salesforce/Lead.csv")
Account = sc.textFile("s3n://development.source.geeksinsight/Datalake/Salesforce/Account.csv")

parts = Task.map(lambda l: l.split(","))
task = parts.map(lambda p: Row(year=p[0],cid=p[1],isbn=p[2],seller=p[3],price=int(p[4])))
schematask = sqlContext.createDataFrame(task)
schematask.registerTempTable("task")


parts2 = Lead.map(lambda l: l.split(","))
task = parts2.map(lambda p: Row(cid=p[0],name=p[1],age=p[2],city=p[3],sex=p[4]))
schemalead = sqlContext.createDataFrame(lead)
schemalead.registerTempTable("lead")

parts3 = Account.map(lambda l: l.split(","))
account = parts3.map(lambda p: Row(isbn=p[0],name=p[1]))
schemaaccount = sqlContext.createDataFrame(account)
schemaaccount.registerTempTable("account")

result_query = sqlContext.sql("""select status, FROM task t left join account a on t.WhoId=a.Id left join lead l on t.WhoId=l.id """)

result = result_query.rdd.map(lambda p: "status: " + p.status).collect()
rdd.saveAsTextFile("output_directory")

for status in result:
    print(status)

#df.write().parquet("s3n://development.source.geeksinsight/Datalake/Salesforce/result.txt")
#df.write.mode("append").format("csv").save("s3n://development.source.geeksinsight/Datalake/Salesforce/result.txt");
```



he first query selects rows from a `.csv` text file. The file contains seven records:

```
$ more plays.csv

1599,As You Like It
1601,Twelfth Night
1594,Comedy of Errors
1595,Romeo and Juliet
1596,The Merchant of Venice
1610,The Tempest
1599,Hamlet
```

Drill recognizes each row as an array of values and returns one column for each row.

```
0: jdbc:drill:zk=local> select * from dfs.`/Users/brumsby/drill/plays.csv`;

+------------------------------------+
|               columns              |
+------------------------------------+
| ["1599","As You Like It"]          |
| ["1601","Twelfth Night"]           |
| ["1594","Comedy of Errors"]        |
| ["1595","Romeo and Juliet"]        |
| ["1596","The Merchant of Venice"]  |
| ["1610","The Tempest"]             |
| ["1599","Hamlet"]                  |
+------------------------------------+
7 rows selected (0.089 seconds)
```

Key obstacles in achieving datalake goals

Technology Challenges

- Data Science model required data scientists who understand different sets of data

Data Swamps

- Data become swamps if not organised properly, over the time the raw data becomes useless if not discovered properly

Organisational Challenges

- Adoption of new model , to do things differently, coming out of traditional ETL approach

Business Value

- Datalake platform is an asset to organisation to do adhoc or on demand analysis on data. So business value will be derived only with definite goals

Data Discovery

- Developers/Scientists/Business Analysts will need more time for Data Discovery and understand the patterns

Summary

Datalake's are needed to solve the complex analytical problems

Datalake's are not replacement of datawarehouses

Datalake are generally built on Hadoop Filesystem

Datalake can be built on Object Store

Almost all Popular vendors has similar patterns of Datalake

Datalake is not a technology, it's a combination of multiple technologies

To make right use of Data in the Datalake , we must have right platform + right tools+ right interface

GeeksInsight You Tube Channel for Videos in this Presentation

Topic	URL
Oracle Object Storate	https://youtu.be/L6Tz66J77G8
Oracle Event HubService	https://youtu.be/YRQWxXZNLzY
Oracle Event Hub Kakfa Topic	https://youtu.be/5mMx4CFBeVE
Oracle DB Cloud	https://youtu.be/jTSdra5uEK4
Oracle Big Data Cloud	https://youtu.be/zRcJFFLjqAc
Oracle Data Integration Platform	https://youtu.be/oAydpTAu_k4
Oracle Big Data Discovery in Datalake	https://youtu.be/WV7s4NC7nWc
Autonomous Datawarehouse Provisioning	https://youtu.be/JY9nhTCuEzg
Autonomous Datawarehouse copy to datalake	https://youtu.be/_HeB-w7w8cY

Q & A

Thanks for joining…

You can reach me at

Blog: db@geeksinsight.com

Linkedin: https://www.linkedin.com/in/suresh-gandhip/

Twitter: https://twitter.com/geeksinsights

@geeksinsights