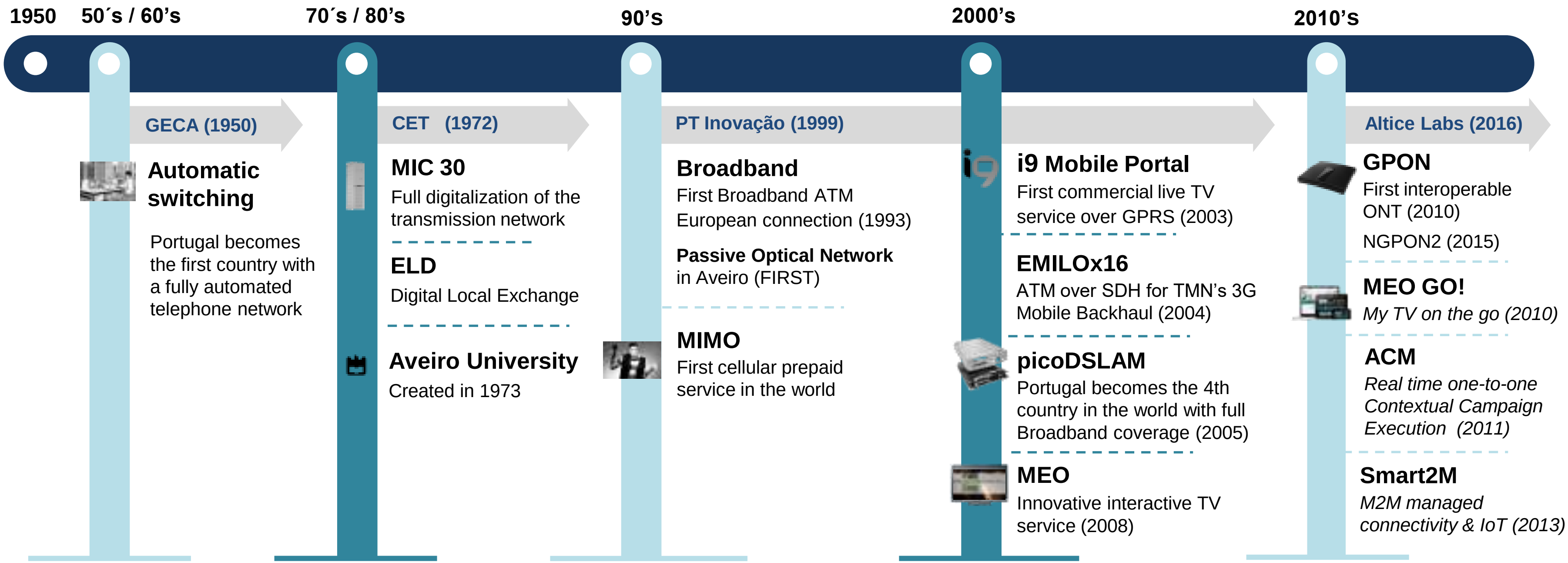
Big Data platform architecture at Altice Labs

Mário Moreira
Head of Experimentation and Technology Coordination

DSPT
DATA SCIENCE PORTUGAL

altice
labs

The history of Altice Labs is linked to the Portuguese telecommunications sector evolution



1950

50´s / 60's

GECA (1950)

Automatic switching

Portugal becomes the first country with a fully automated telephone network

- *Analogue Switching*

70´s / 80's

CET (1972)

MIC 30

Full digitalization of the transmission network

ELD

Digital Local Exchange

Aveiro University

Created in 1973

- *Electronics*
- *Digital Transmission and Switching*
- *Network Management Systems*

90's

PT Inovação (1999)

Broadband

First Broadband ATM European connection (1993)

Passive Optical Network in Aveiro (FIRST)

MIMO

First cellular prepaid service in the world

- *International R&D collaboration*
- *Broadband Transmission*
- *Intelligent Networks*
- *Television (IPTV)*

2000's

i9 Mobile Portal

First commercial live TV service over GPRS (2003)

EMILOx16

ATM over SDH for TMN's 3G Mobile Backhaul (2004)

picoDSLAM

Portugal becomes the 4th country in the world with full Broadband coverage (2005)

MEO

Innovative interactive TV service (2008)

- *Internationalization*
- *Mobile Networks*
- *Multimedia and Television*

2010's

Altice Labs (2016)

GPON

First interoperable ONT (2010)

NGPON2 (2015)

MEO GO!

My TV on the go (2010)

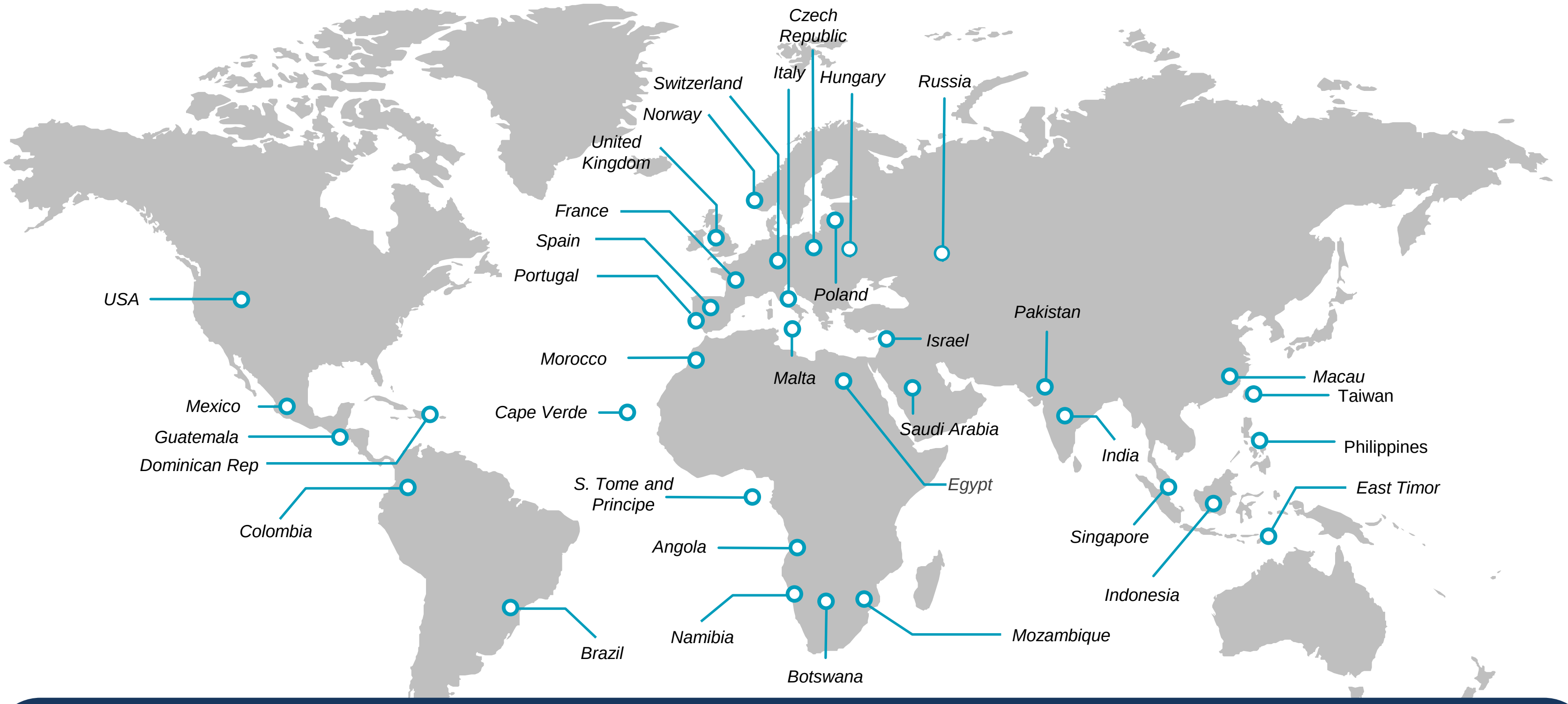
ACM

Real time one-to-one Contextual Campaign Execution (2011)

Smart2M

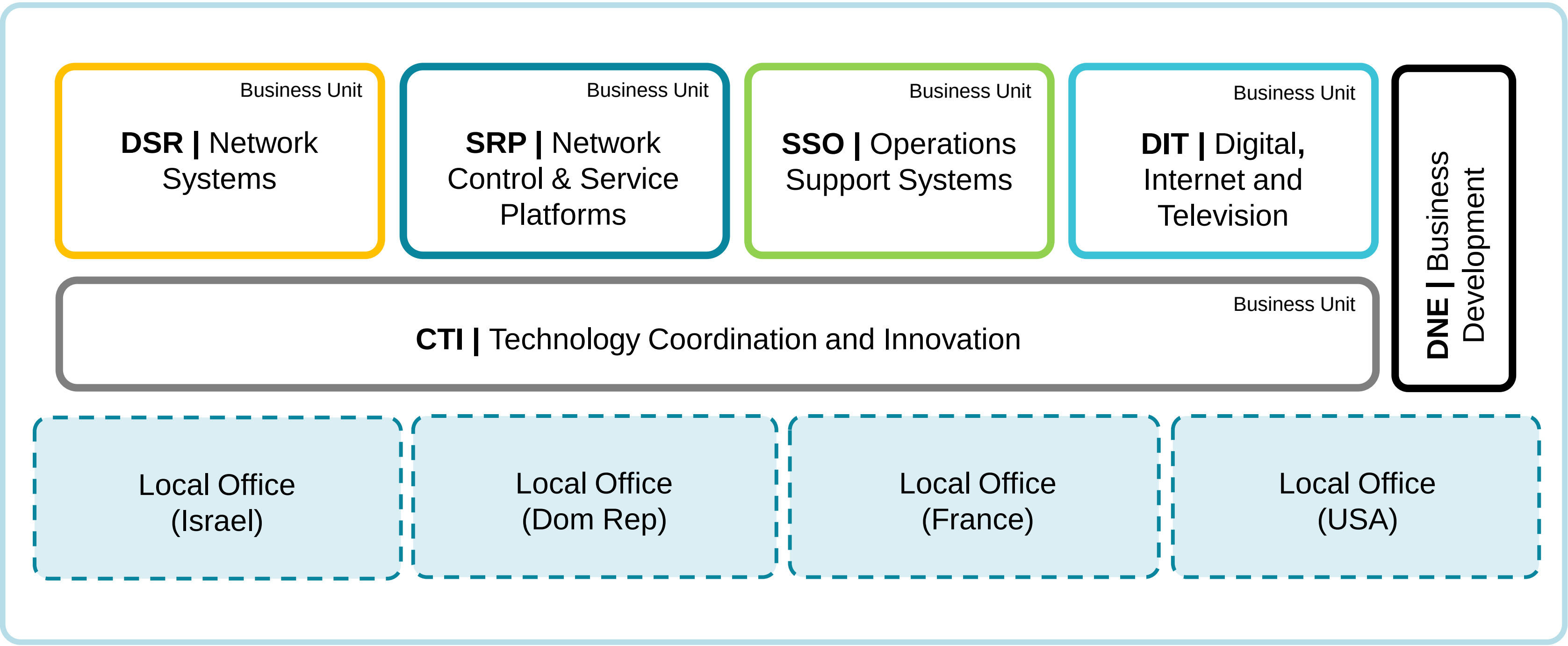
M2M managed connectivity & IoT (2013)

- *Optical access*
- *Mobility*
- *Digitalization/Virtualization*
- *Personalization / Contextualization*
- *M2M and IoT*



200 million people communicate everyday through technology developed by Altice Labs

Governance



Business Unit

DSR | Network Systems

Business Unit

SRP | Network Control & Service Platforms

Business Unit

SSO | Operations Support Systems

Business Unit

DIT | Digital, Internet and Television

DNE | Business Development

Business Unit

CTI | Technology Coordination and Innovation

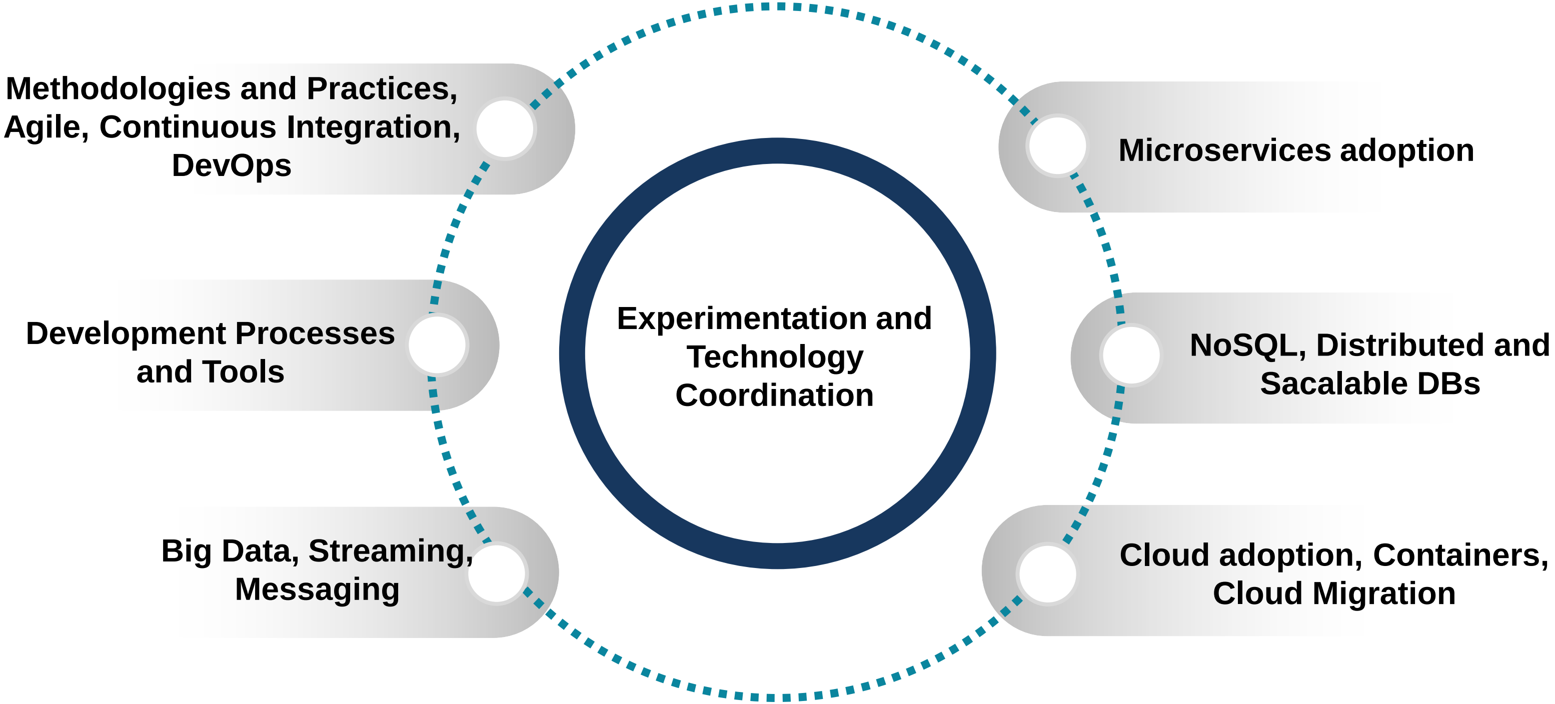
Local Office (Israel)

Local Office (Dom Rep)

Local Office (France)

Local Office (USA)

Current Innovation Areas



Experimentation and Technology Coordination

Methodologies and Practices, Agile, Continuous Integration, DevOps

Development Processes and Tools

Big Data, Streaming, Messaging

Microservices adoption

NoSQL, Distributed and Sacalable DBs

Cloud adoption, Containers, Cloud Migration

Big Data @ Altice Labs

Big Data @ Altice Labs

4T 2012

Phase 1

To experiment and to obtain know how in Big Data Technologies

Technologies used Cloudera Hadoop, Map Reduce, Pig, Impala, Hbase, Flume, Sqoop, Oozie, Zookeeper, …

Benchmarks against the traditional approach (Oracle DB + Java code)

Learn what works and what doesn’t work

2013 - 2014

Phase 2

To “sell” and disseminate Big Data technologies to product teams

Create a reference Big Data technology stack and a technology adoption strategy for ALB products

Comparative studies on NoSQL Databases (MongoDB, Cassandra, Redis, Neo4j, VoltDB, HBase)

2015

Phase 3

Experiment with Real-time Big Data technologies

Select and validate the technologies better suited to implement high performance and high scalability versions of some ALB products

Comparative analysis and benchmark of Storm and Spark (we didn’t consider Flink and Samza mature enough at that time)

2016 - Today

Phase 4

Management of MEO Big Data platform (ex-Sapo)

Most data is from Web Portal, mobile apps and IPTV

Participation on Altice Big Data Group project to standardize Big Data technology stack, reference architecture and data models and governance

About the “Golias” platform

NUMBERS

Total of 98 servers holding 191 TB (compressed data) and 8 TB (uncompressed aggregated data) divided by 10 different clusters

Data is received from +40 different distinct systems and is processed by +240 distinct processes

SOFTWARE USED

Cloudera for the Hadoop clusters, 3 PostgreSQL clusters, 2 Cassandra clusters, MonetDB, and 49 instances of Solr

Data is handled by Pig, Java or Python processes

DATA INGESTION

By batch processes that import data from external databases

By Sapo Broker, and in house developed pub/sub messaging platform

By service listeners that receive events from webpages, mobile apps, set top boxes apps, etc

DATA ACCESS

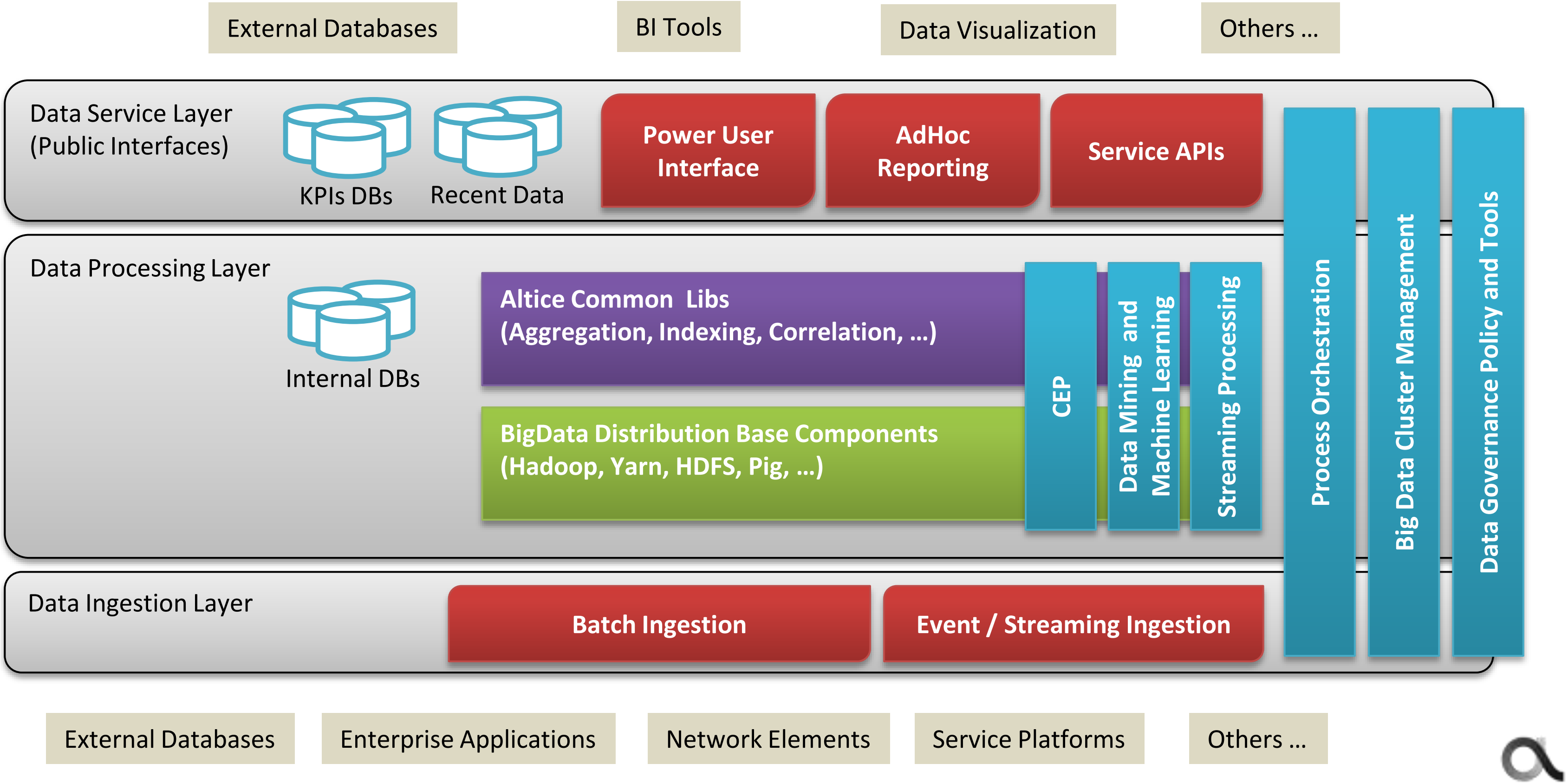
In house developed web applications for analytics and dashboards

Using in house developed API interfaces

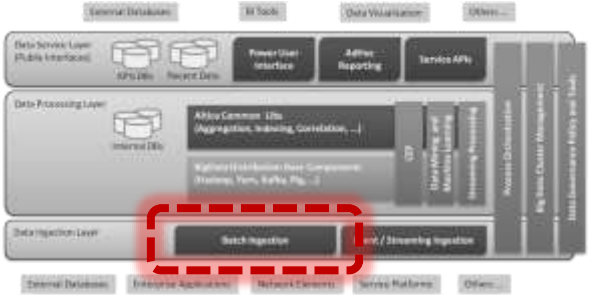
Ad hoc reports as needed by the business areas

Altice Labs Big Data reference architecture

Reference Architecture



Data Ingestion - Batch



1 Explore

- Apache NiFi
- StreamSets

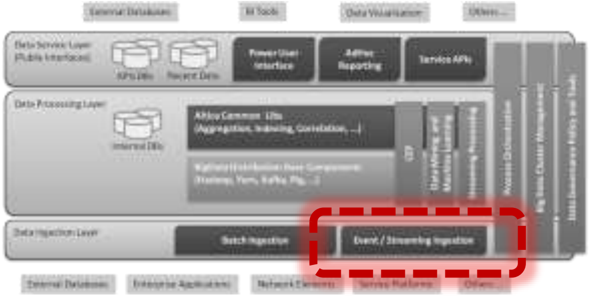
2 Experiment

- Sqoop
- Flume

3 Adopt

- Shell scripts
- Python

Data Ingestion - Streaming



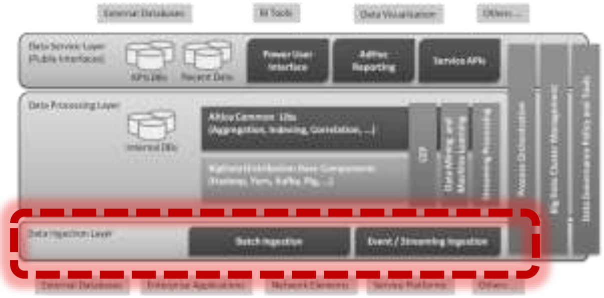
1 Explore

2 Experiment

3 Adopt

- Kafka
- REST APIs

Data Ingestion – Shared Storage



1 Explore

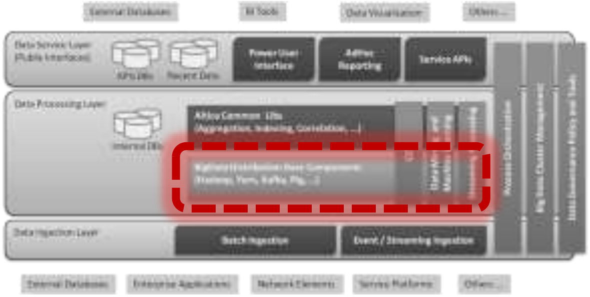
- CephFS
- BeeGFS (old FGFS from Fraunhofer)

2 Experiment

3 Adopt

- GlusterFS

Hadoop Base Distribution



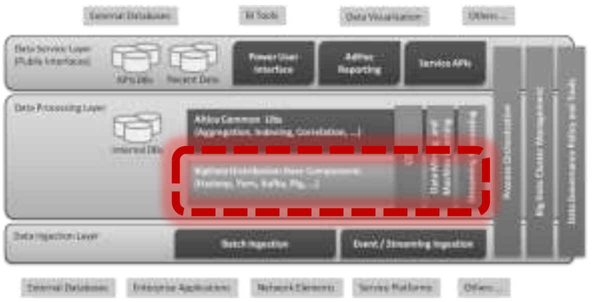
① **Explore**

② **Experiment**

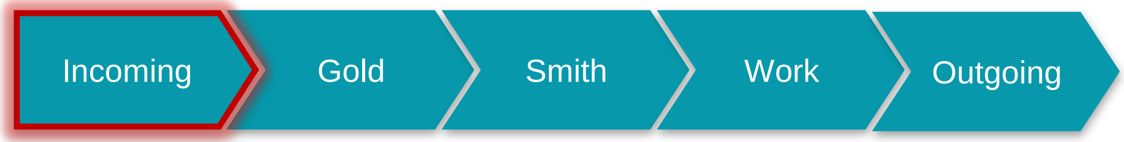
③ **Adopt**

- Cloudera

Data Layer – Incoming stage



Data Stages

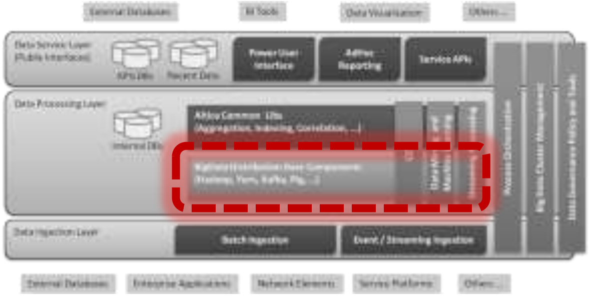


1 Explore

2 Experiment

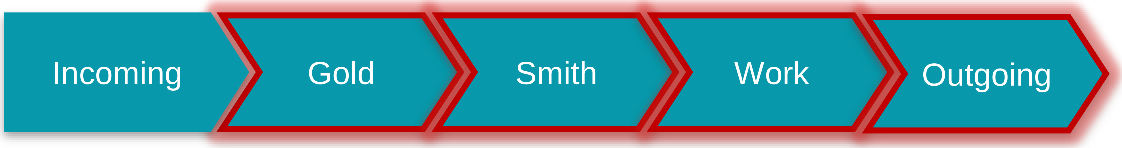
3 Adopt

- Kudu
- HDFS



Data Layer – Gold and other stages

Data Stages



Incoming

Gold

Smith

Work

Outgoing

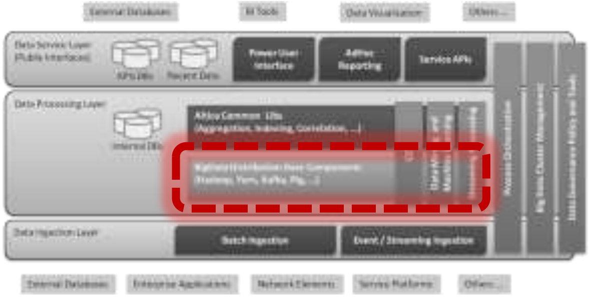
1 Explore

2 Experiment

3 Adopt

- Impala
- HDFS

Data Processing - Batch



1 Explore

2 Experiment

3 Adopt

- PIG
- Python

Data Processing - Streaming

External Databases
BI Tools
Data Visualisation
Others...
Data Service Layer
(Public Interfaces)
KPIs DBs
Recent Data
Power User Interface
Altice Reporting
Service APIs
Data Processing Layer
Internal DBs
Altice Common Libs
(Aggregation, Indexing, Correlation, ...)
Hadoop Distribution Base Components
(Hadoop, Yarn, Kafka, Pig, ...)
ETL
Big Data Cluster Management
Data Ingestion Layer
Batch Ingestion
Event / Streaming Ingestion
External Databases
Enterprise Applications
Network Elements
Service Platforms
Others...

1 Explore

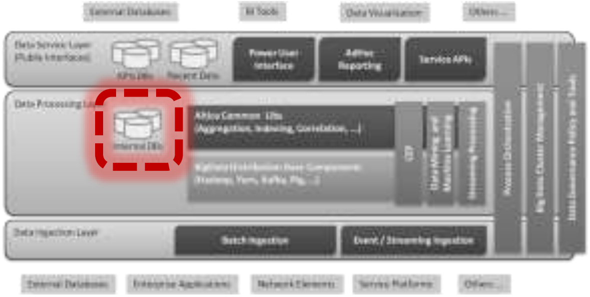
- Spark

2 Experiment

- Storm
- Flink

3 Adopt

Working In-Memory Storage



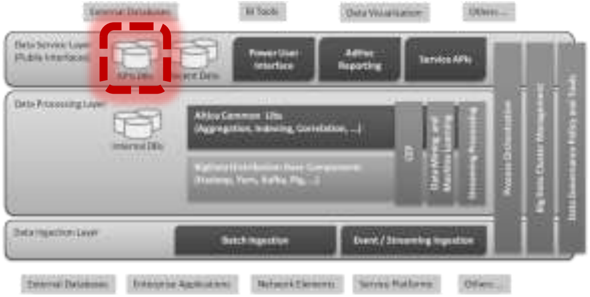
1 Explore

2 Experiment

3 Adopt

- Redis

Data Warehouse



1 Explore

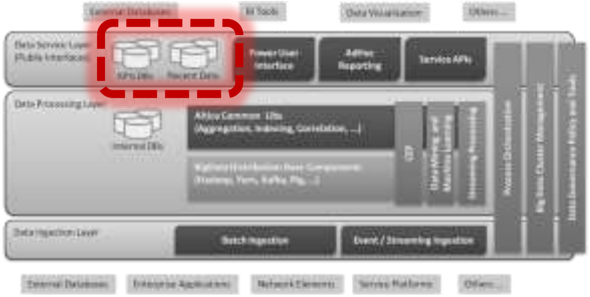
2 Experiment

- Citus DB
- Apache Drill

3 Adopt

- PostgreSQL

Data Science Playground - Storage



1 Explore

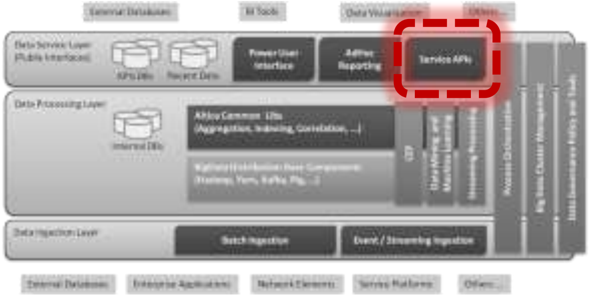
2 Experiment

- Citus DB
- Apache Drill

3 Adopt

- HDFS
- Impala
- PostgreSQL

Data Access APIs



1 Explore

2 Experiment

- Apache Drill
- REST APIs

3 Adopt

Cluster Management and Monitoring

External Databases
BI Tools
Data Visualisation
Others...
Data Service Layer
(Public Interfaces)
KPIs DBs
Recent Data
Power User Interface
Altice Reporting
Service APIs
Data Processing Layer
Internal DBs
Altice Common Libs
(Aggregation, Indexing, Correlation, ...)
Hadoop Distribution Base Components
(Hadoop, Yarn, Kafka, Pig, ...)
CEP
Data Mining and Machine Learning
Streaming Processing
Big Data Cluster Management
Data Ingestion Layer
Batch Ingestion
Event / Streaming Ingestion
External Databases
Enterprise Applications
Network Elements
Service Platforms
Others...

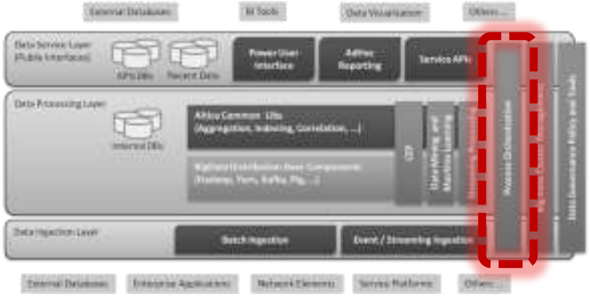
1 Explore

2 Experiment

3 Adopt

- Cloudera Manager

Process Orchestration



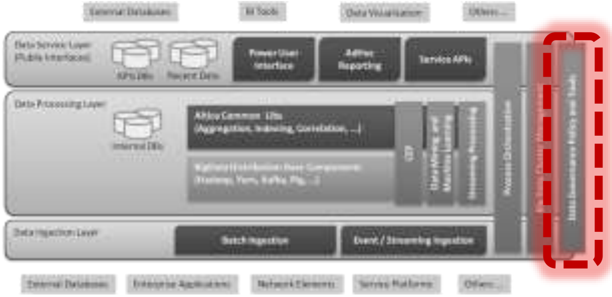
1 Explore

2 Experiment

- Apache Airflow

3 Adopt

Data Governance, Lineage and Audit



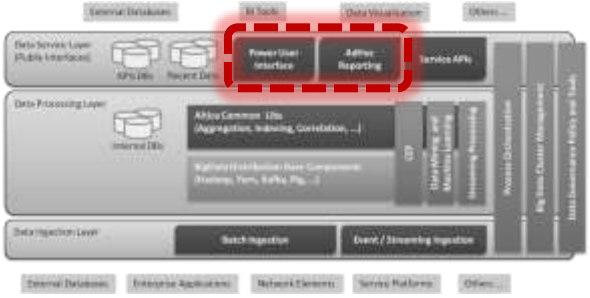
1 Explore

- Cloudera Navigator
- Apache Atlas

2 Experiment

3 Adopt

Power User Interface and Ad Hoc Reporting



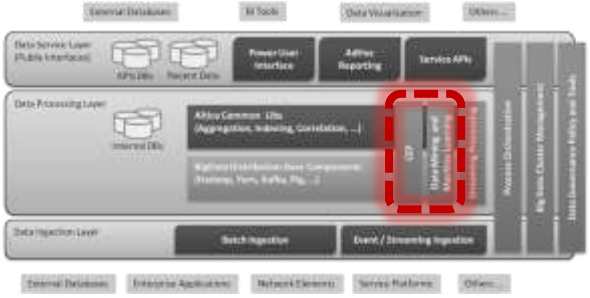
1 Explore

2 Experiment

- Apache Zepplin
- Jupyter

3 Adopt

CEP and Machine Learning



1 Explore

- Esper (CEP) (within storm)
- Apache Mahout
- Python – numpy, pandas, scikit-learn
- Yet no experience with Tensorflow, Mxnet, H2O, Theano, Torch, …

2 Experiment

3 Adopt

About the Technologies Presented

Data visualization and exploration tools are out of this scope (assuming reuse of already installed tools like Tableau and others).

Bussiness users need to be able to take advantage of the Data Lake. Traditional BI and Analytical tools may not play well with Hadoop ecosystem of technologies (proprietary formats, informaton silos, designed assuming relational model, …). We will need new tools that integrate well with these technologies for bussiness users (ex: Dataiku, others). This might imply user retraining.

It’s still a Work in Progress – as we experiment and learn more we might change a few things.

Your comments and sugestions are wellcome.

Big Data platform architecture at Altice Labs

Mário Moreira
Head of Experimentation and Technology Coordination

Any questions?