ITBA Escuela de INNOVACIÓN

EVOLUCIÓN CONTINUA

Especialización en Ciencia de Datos

Herramientas para Grandes Volúmenes de Datos

Data Lake Tools

www.itba.edu.ar

Agenda de Hoy

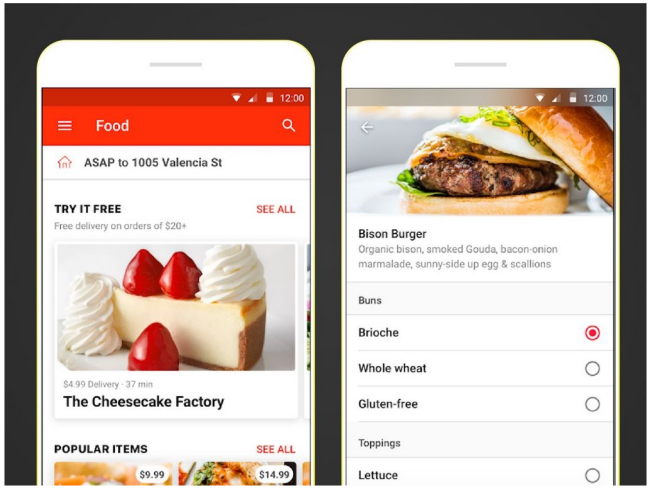
01 Introducción y blueprints

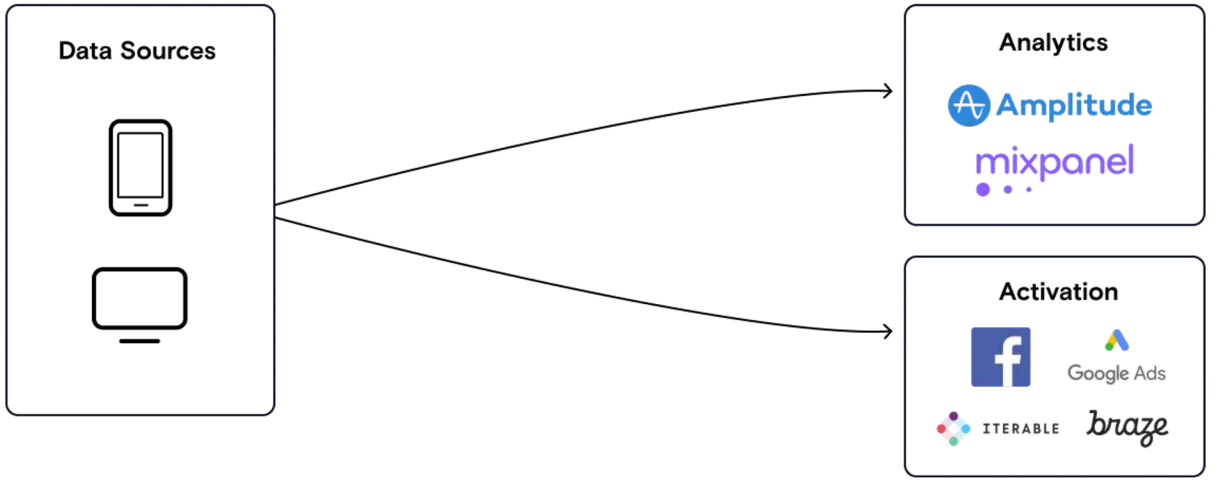
02 Componentes y herramientas del Data Lake

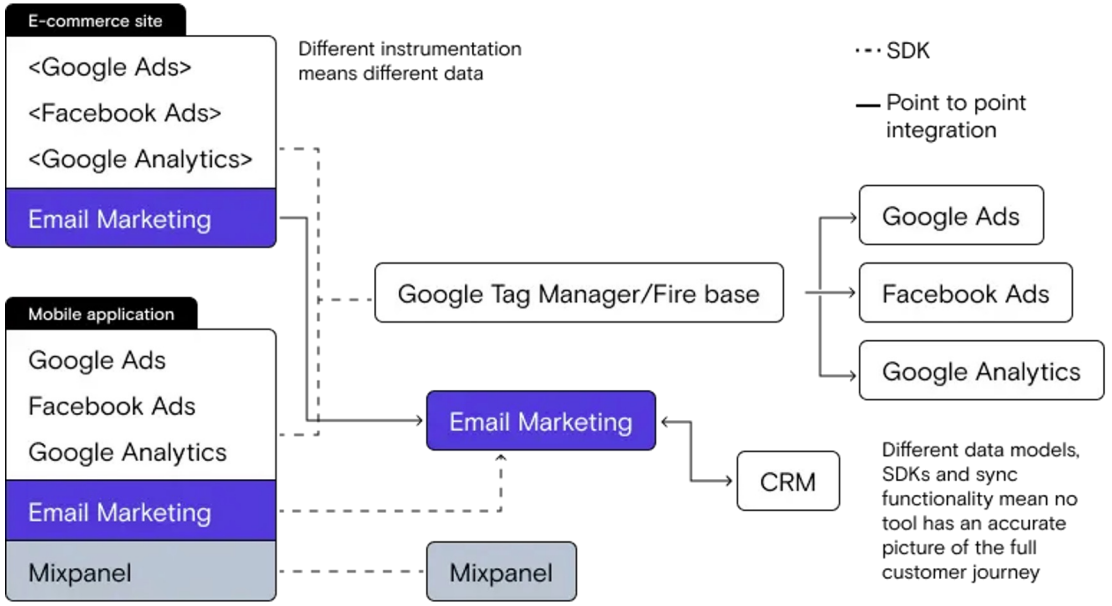
03 Arquitecturas delta y data quality

04 Práctica

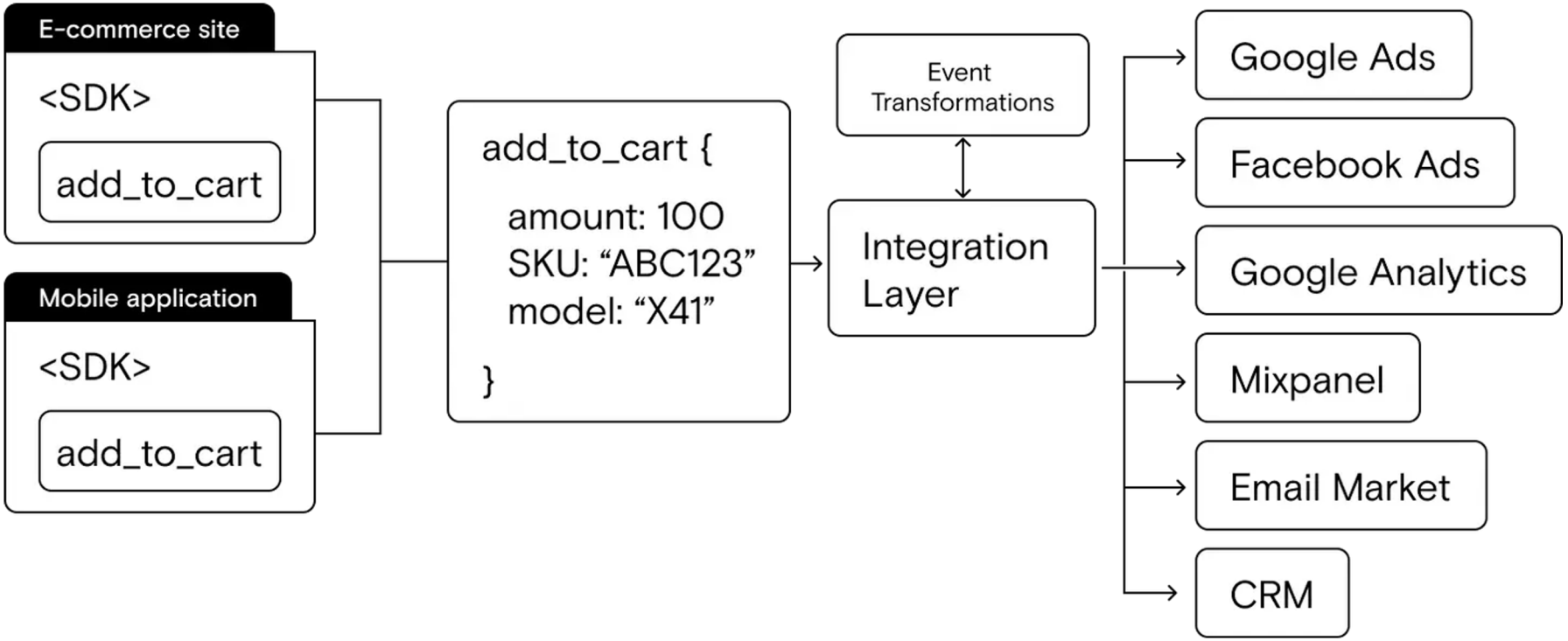
01 | Introducción y blueprints


12:00
Food
ASAP to 1005 Valencia St
TRY IT FREE
SEE ALL
Free delivery on orders of $20+
$4.99 Delivery · 37 min
The Cheesecake Factory
POPULAR ITEMS
SEE ALL
$9.99
$14.99
12:00
Bison Burger
Organic bison, smoked Gouda, bacon-onion marmalade, sunny-side up egg & scallions
Buns
Brioche
Whole wheat
Gluten-free
Toppings
Lettuce

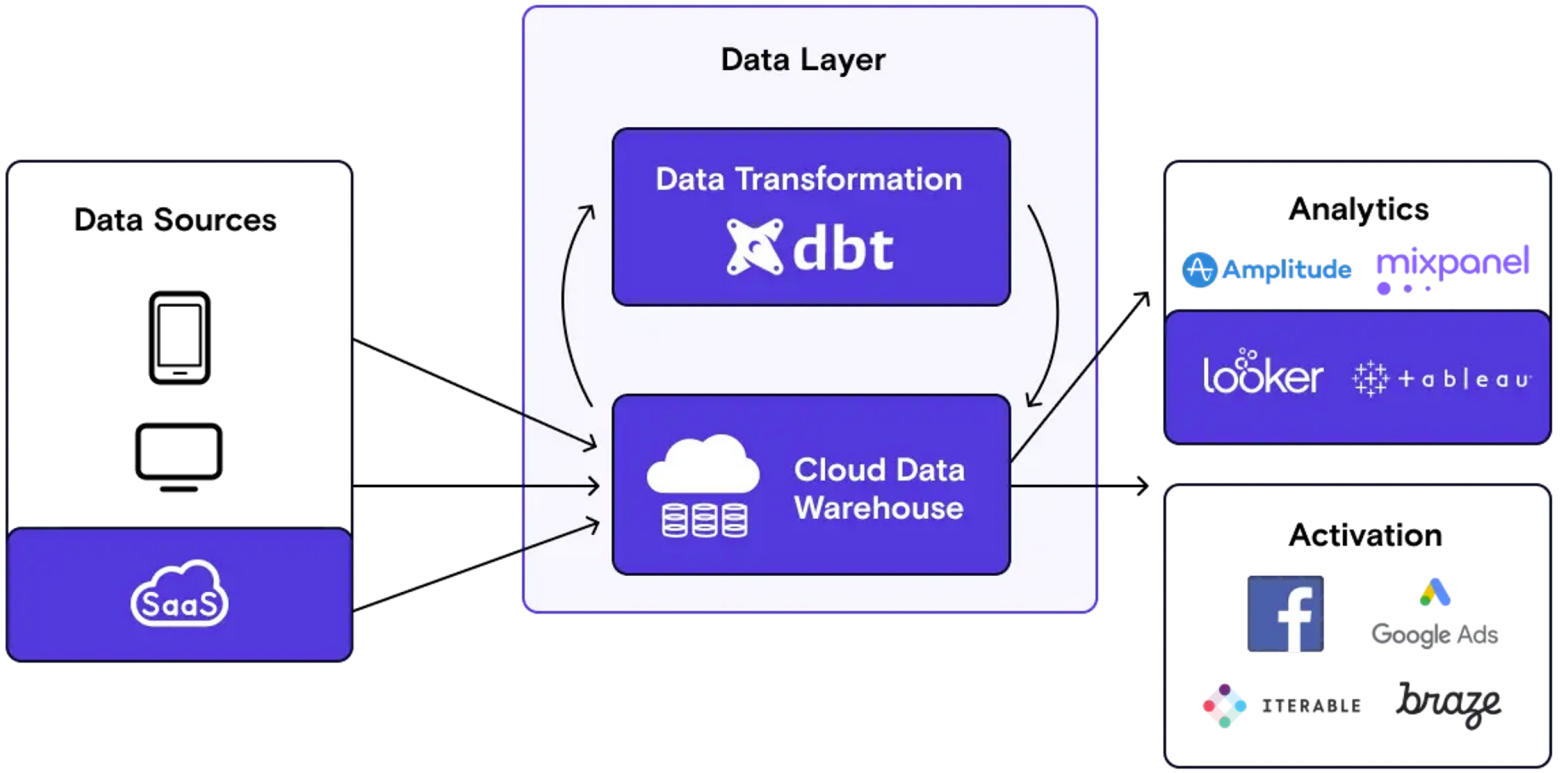

Data Sources
Analytics
Amplitude
mixpanel
Activation
Google Ads
ITERABLE
braze


E-commerce site
<Google Ads>
<Facebook Ads>
<Google Analytics>
Email Marketing
Different instrumentation means different data
SDK
Point to point integration
Google Tag Manager/Fire base
Google Ads
Facebook Ads
Google Analytics
Mobile application
Google Ads
Facebook Ads
Google Analytics
Email Marketing
Mixpanel
Email Marketing
CRM
Mixpanel
Different data models, SDKs and sync functionality mean no tool has an accurate picture of the full customer journey

Integración

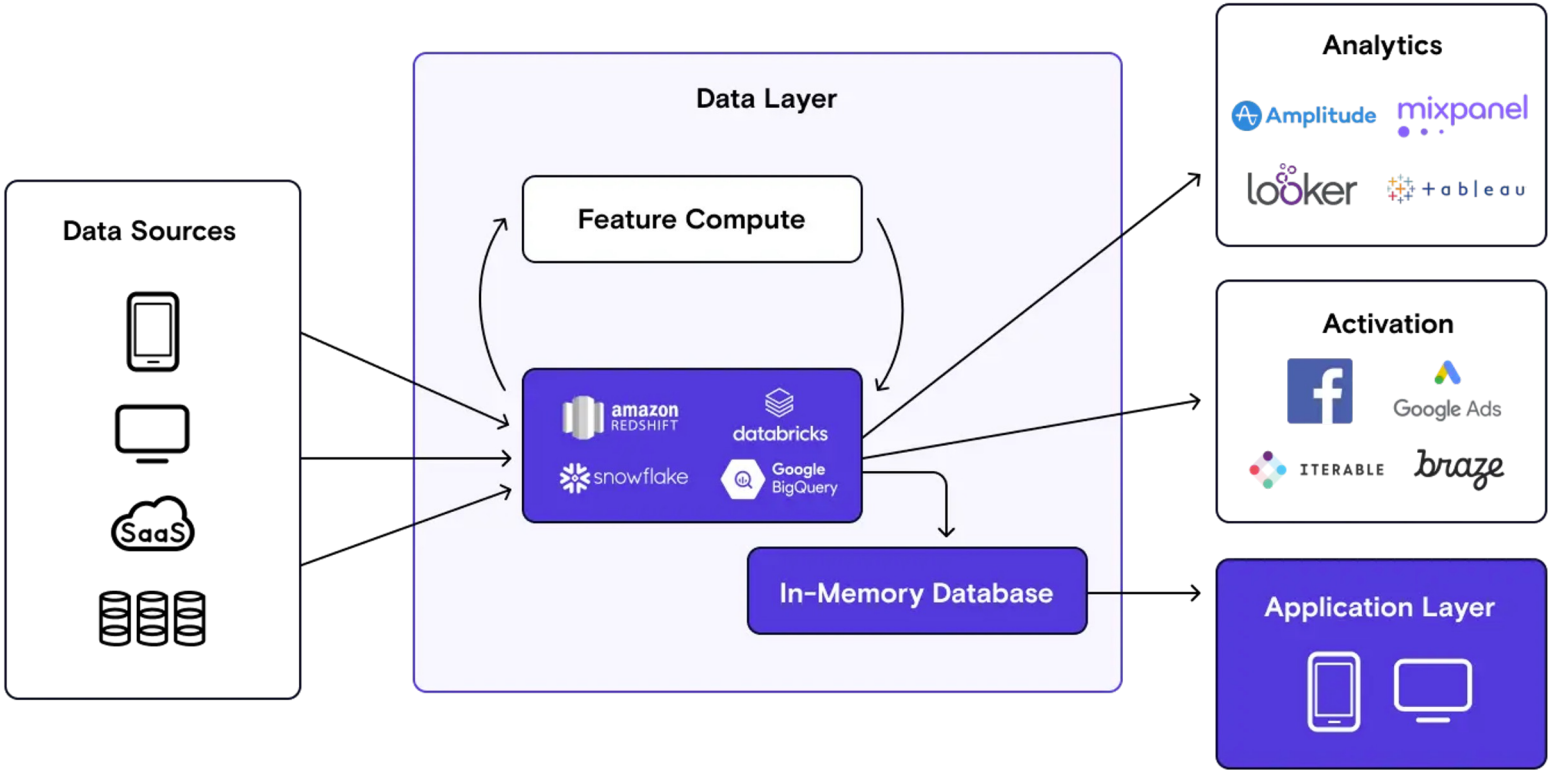


Phase 2: Centralized Data Store

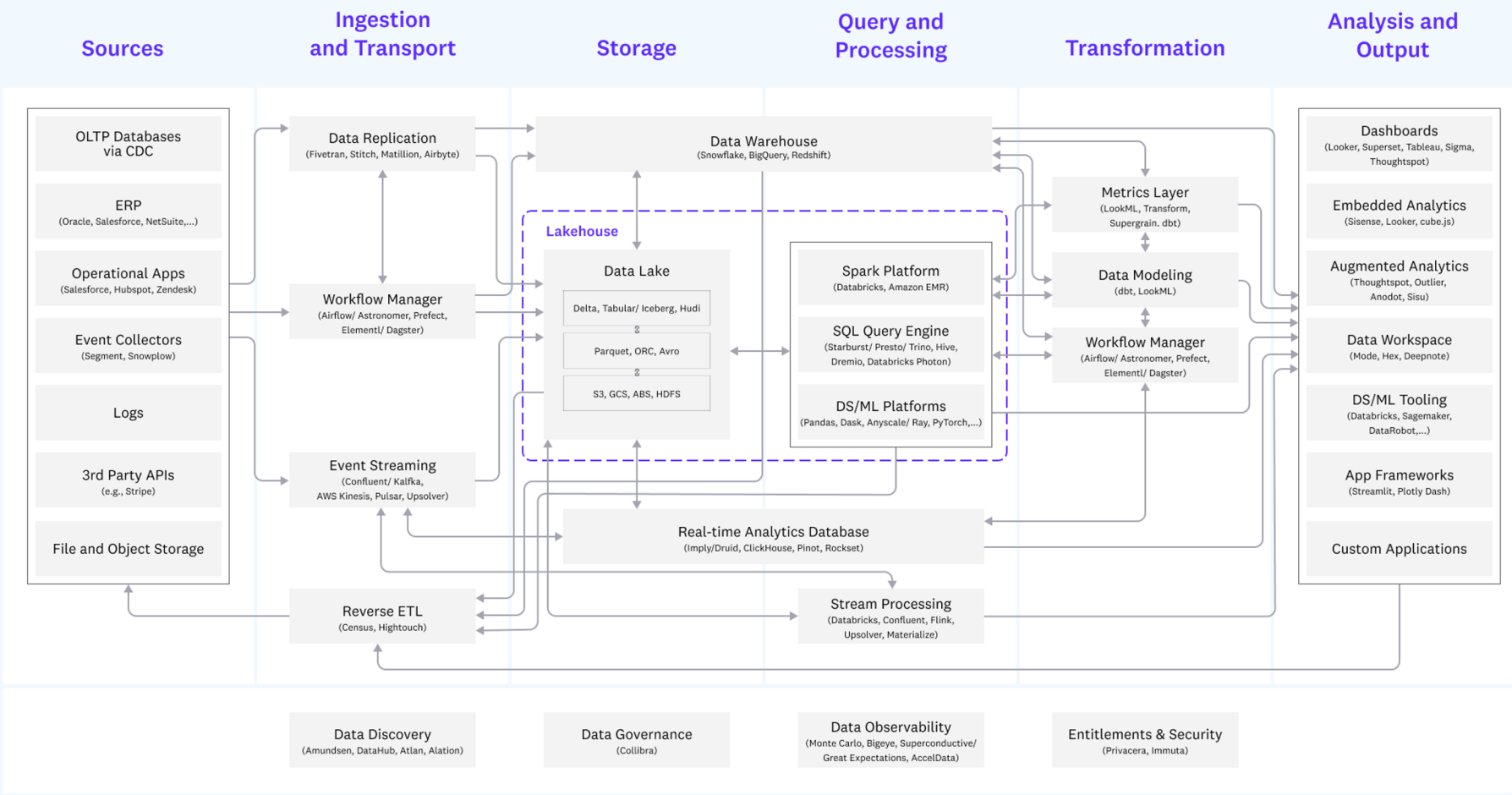


ITBA

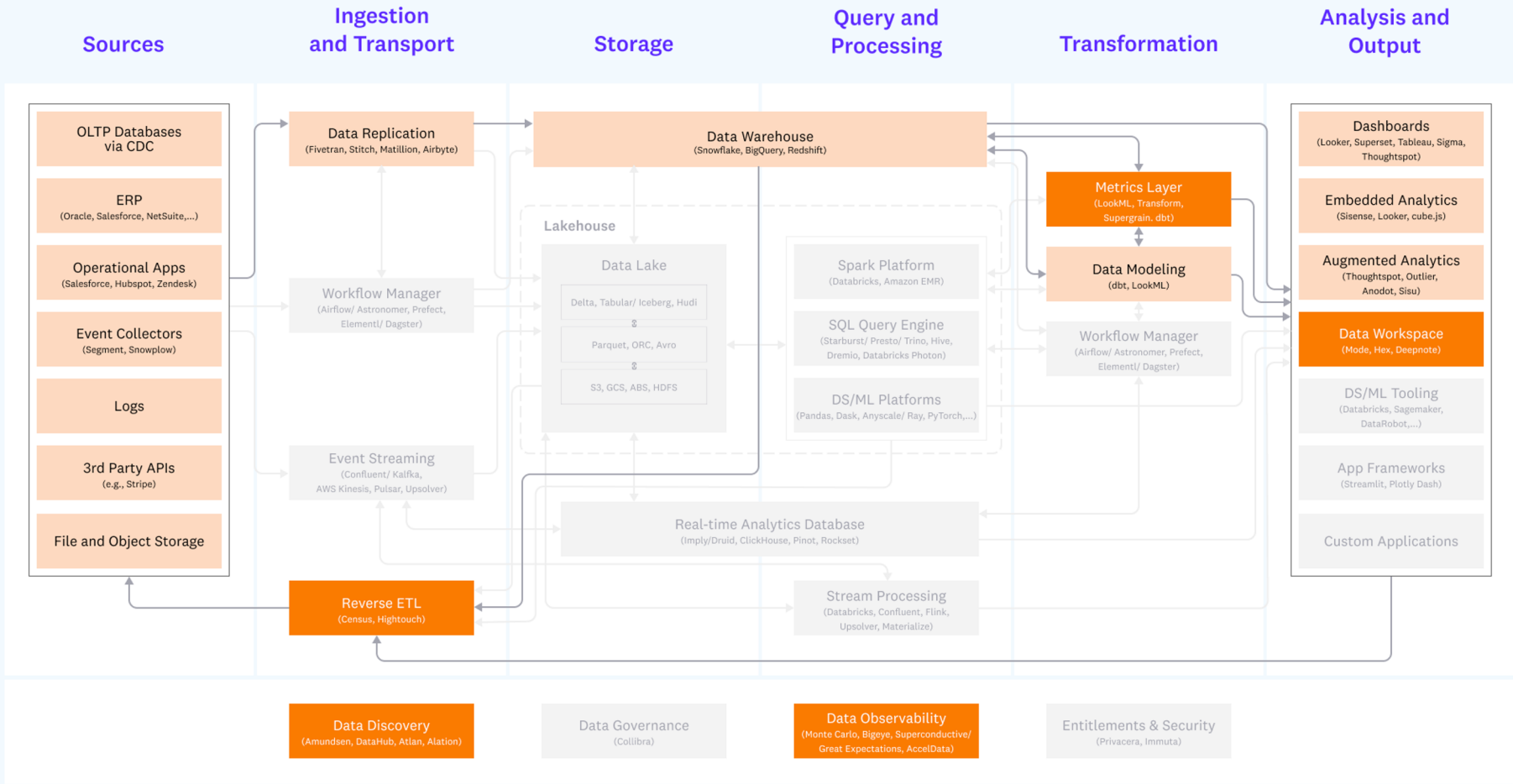
Phase 3: ML and Real-Time



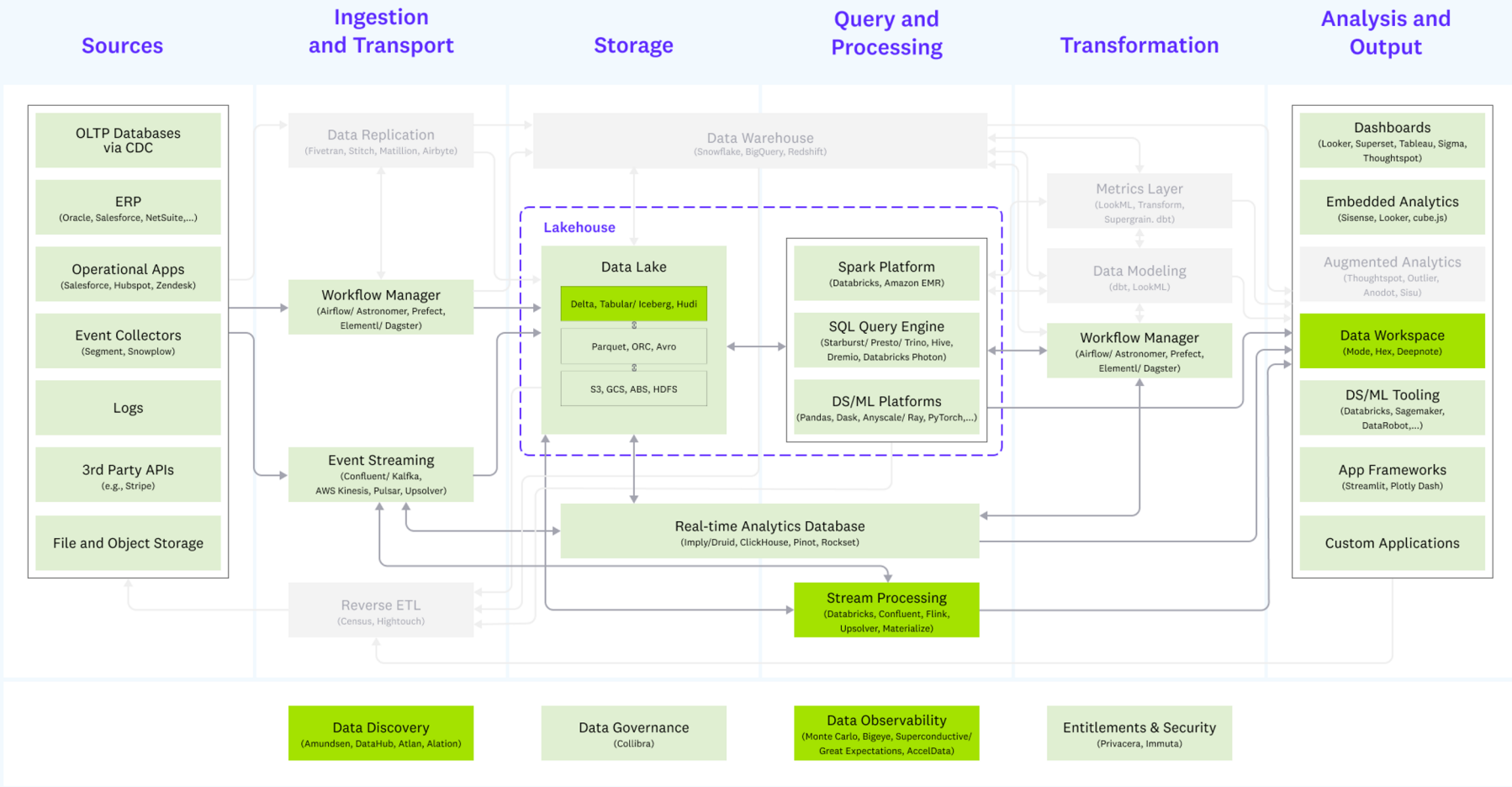
Unified Data Infrastructure (2.0)

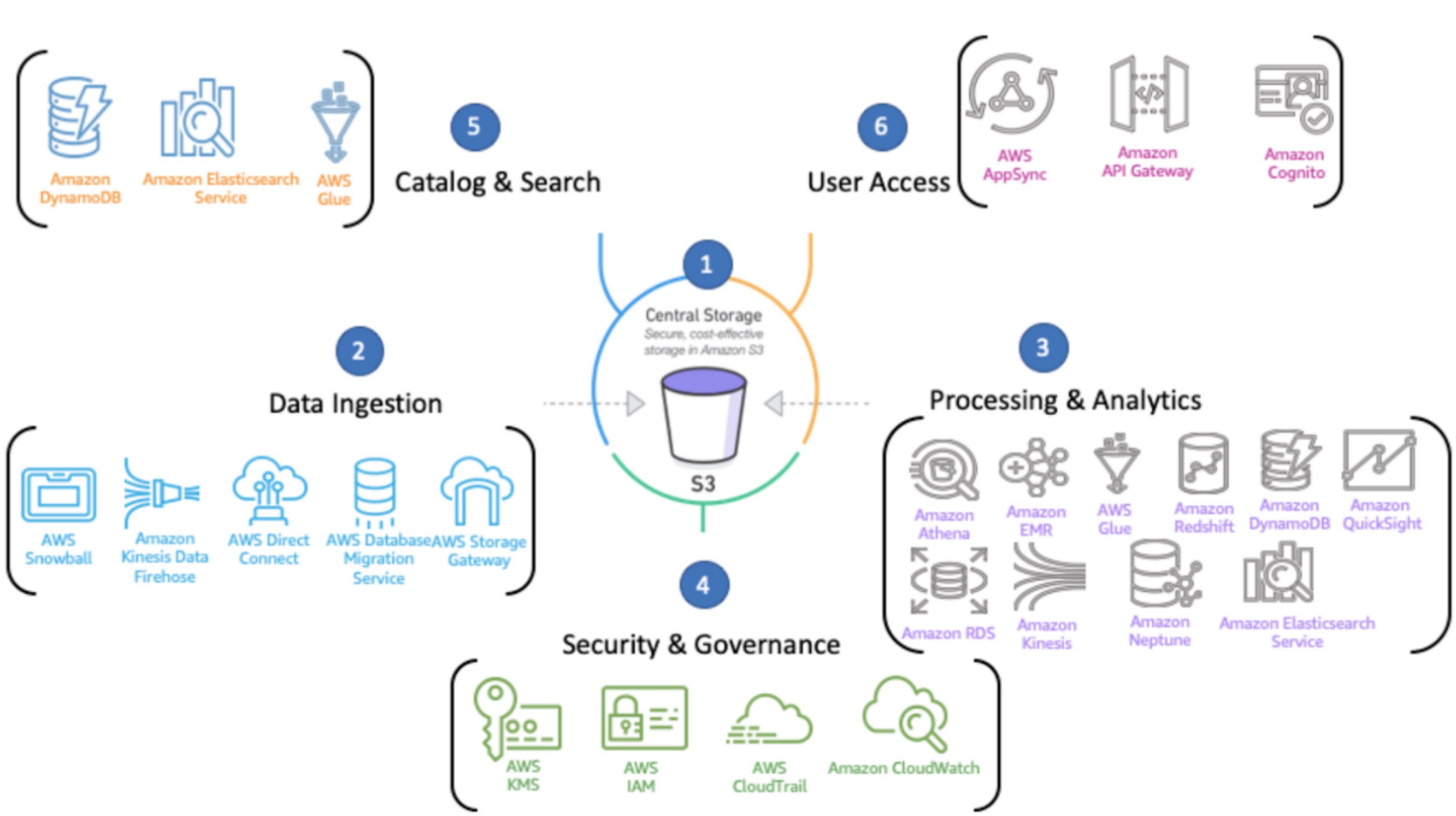


Blueprint 1: Modern Business Intelligence

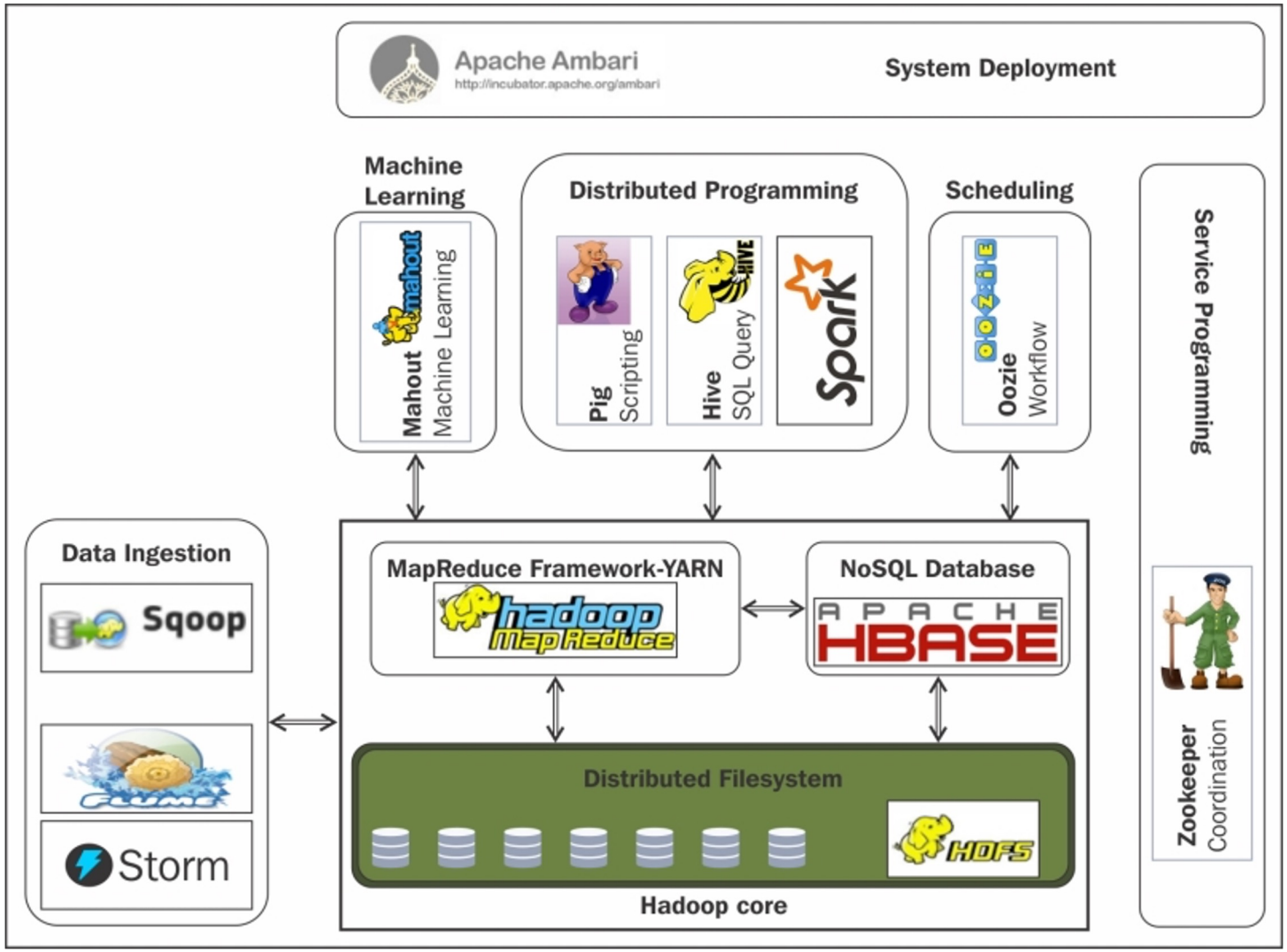


Blueprint 2: Multimodal Data Processing




Amazon DynamoDB
Amazon Elasticsearch Service
AWS Glue
5
Catalog & Search
6
User Access
AWS AppSync
Amazon API Gateway
Amazon Cognito
1
Central Storage
Secure, cost-effective storage in Amazon S3
S3
2
Data Ingestion
AWS Snowball
Amazon Kinesis Data Firehose
AWS Direct Connect
AWS Database Migration Service
AWS Storage Gateway
3
Processing & Analytics
Amazon Athena
Amazon EMR
AWS Glue
Amazon Redshift
Amazon DynamoDB
Amazon QuickSight
Amazon RDS
Amazon Kinesis
Amazon Neptune
Amazon Elasticsearch Service
4
Security & Governance
AWS KMS
AWS IAM
AWS CloudTrail
Amazon CloudWatch

02 | Componentes y herramientas del Data Lake


Apache Ambari
http://incubator.apache.org/ambari
System Deployment
Machine Learning
Mahout Machine Learning
Distributed Programming
Pig Scripting
Hive SQL Query
Spark
Scheduling
Oozie Workflow
Service Programming
Zookeeper Coordination
Data Ingestion
Sqoop
Flume
Storm
MapReduce Framework-YARN
hadoop MapReduce
NoSQL Database
APACHE HBASE
Distributed Filesystem
HDFS
Hadoop core


HIVE

Apache Hive

Apache Hive is a **data warehouse infrastructure built on top of Hadoop** for providing data summarization, query, and analysis.

It was initially developed by Facebook in 2008.

Oriented towards analytical (OLAP) workloads.

In the beginning used MapReduce as backend, now Tez and Spark.

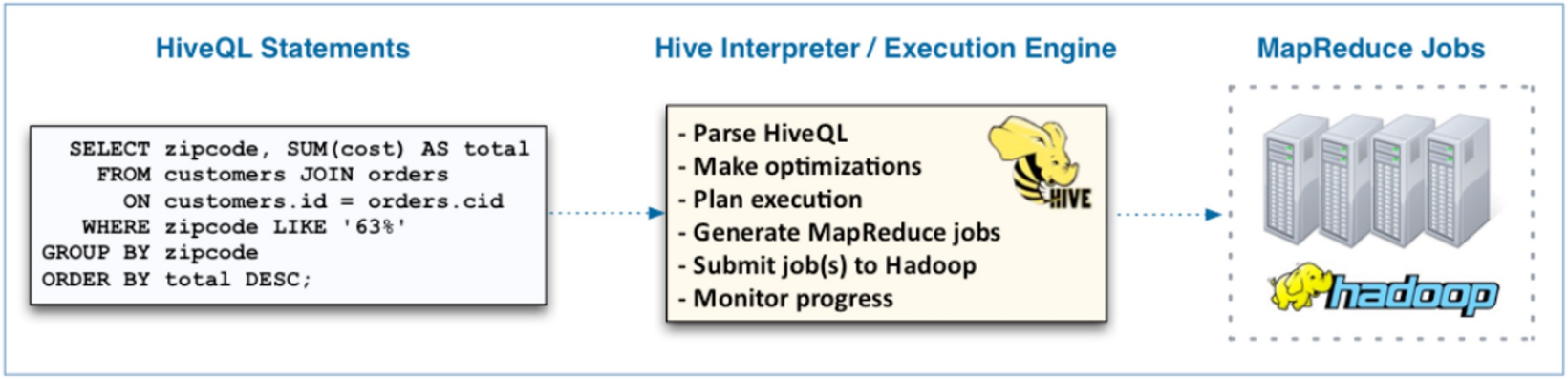
Hive's SQL dialect, called HiveQL, is a mixture of SQL-92, MySQL, Oracle's SQL and SQL:2003.

Apache Hive

Hive runs on the master machine

Converts HiveQL queries to MapReduce jobs.

Submits MapReduce jobs to the cluster.



ITBA

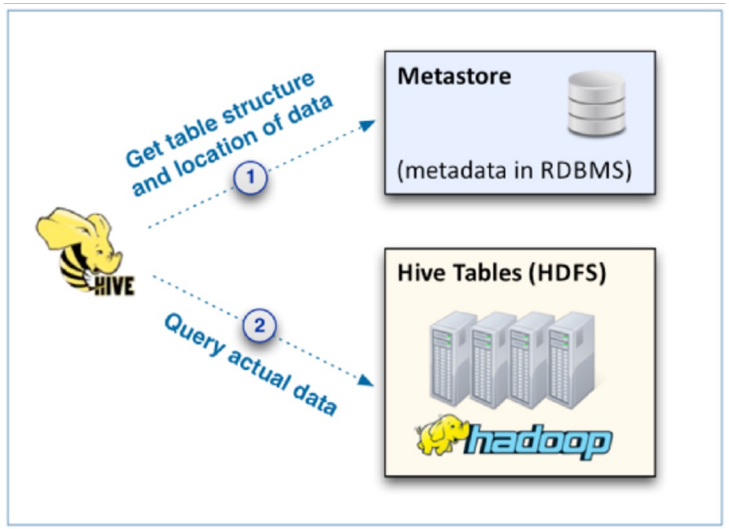
Apache Hive

Hive queries operate on tables; just like any other Data Warehouse.

A table is just an HDFS/S3 directory containing one or more files
Default Path: /user/hive/warehouse/<table_name>

Hive supports many formats for data stored

Hive stores the table metadata (location, fields, format, etc.) in the Metastore a specialized process that uses an RDBMS to store data.



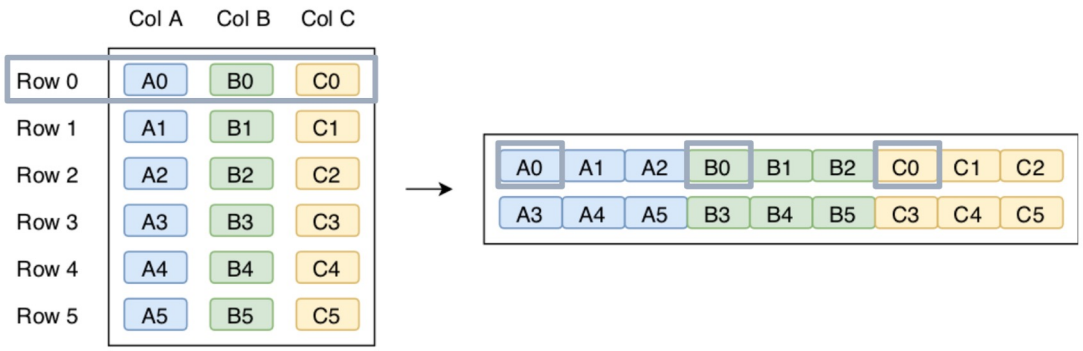
ITBA

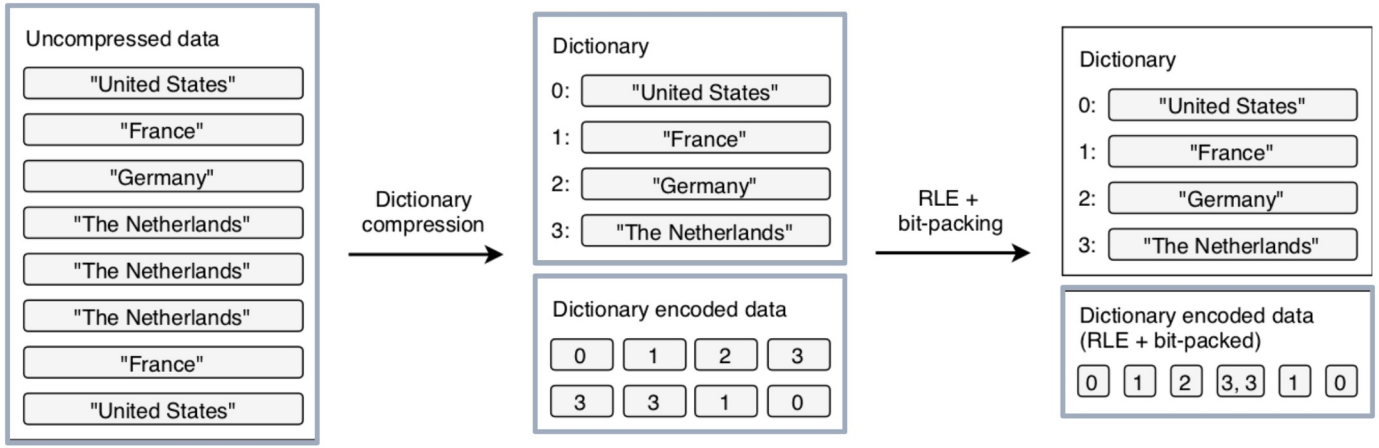
DDL: CREATE TABLE

```
CREATE EXTERNAL TABLE IF NOT EXISTS request_logs (
  timestamp STRING,
  event_type STRING,
  request_ip STRING,
  request_port INT
)
PARTITIONED BY (year STRING, month STRING, day STRING)
  -ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n'
  -ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe' WITH SERDEPROPERTIES ( "input.regex" = "([^ ]*) ([^ ]*) ([^ ]*):([0-9]*)$")
  -STORED AS [TEXTFILE|PARQUET|ORC|RCFILE|AVRO]
LOCATION '[s3|hdfs]://logs.company.com/request_logs/';
```

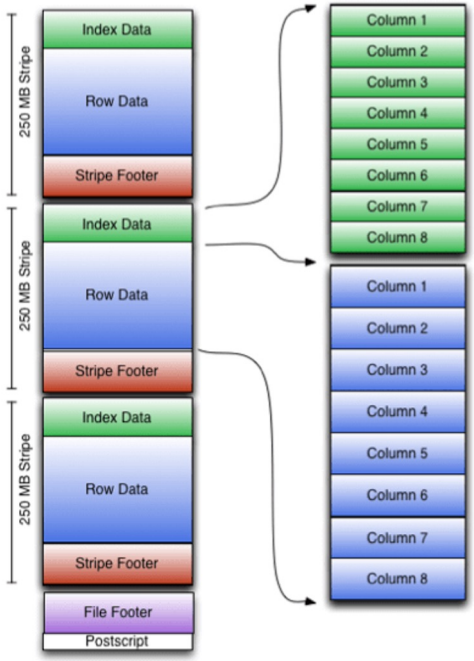
ITBA

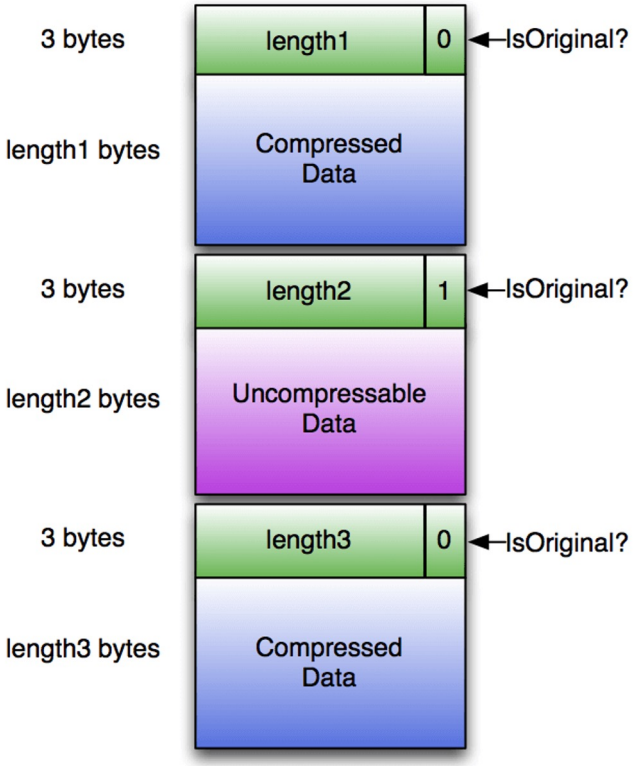
Columnar Format: Parquet





Columnar Format: ORC (Optimized Row Columnar)





DDL: PARTITIONED TABLES

Tables can be partitioned, which is a **physical arrangement of data, into distinct subdirectories for each unique partitioned key**. Partitions can be statically or dynamically added

```
ALTER TABLE request_logs ADD IF NOT EXISTS PARTITION (year = '2021', month = '03', day = '02') LOCATION
  's3://logs.company.com/request_logs/2021/03/02/';
ALTER TABLE request_logs ADD IF NOT EXISTS PARTITION (year = '2021', month = '03', day = '03') LOCATION
  's3://logs.company.com/request_logs/2021/03/03/';
```

If directories where in the format:

's3://logs.company.com/request_logs/year=2021/month=03/day=03/'

We could use the following statement to load dynamically all partitions:

MSCK REPAIR TABLE request_logs;

ITBA

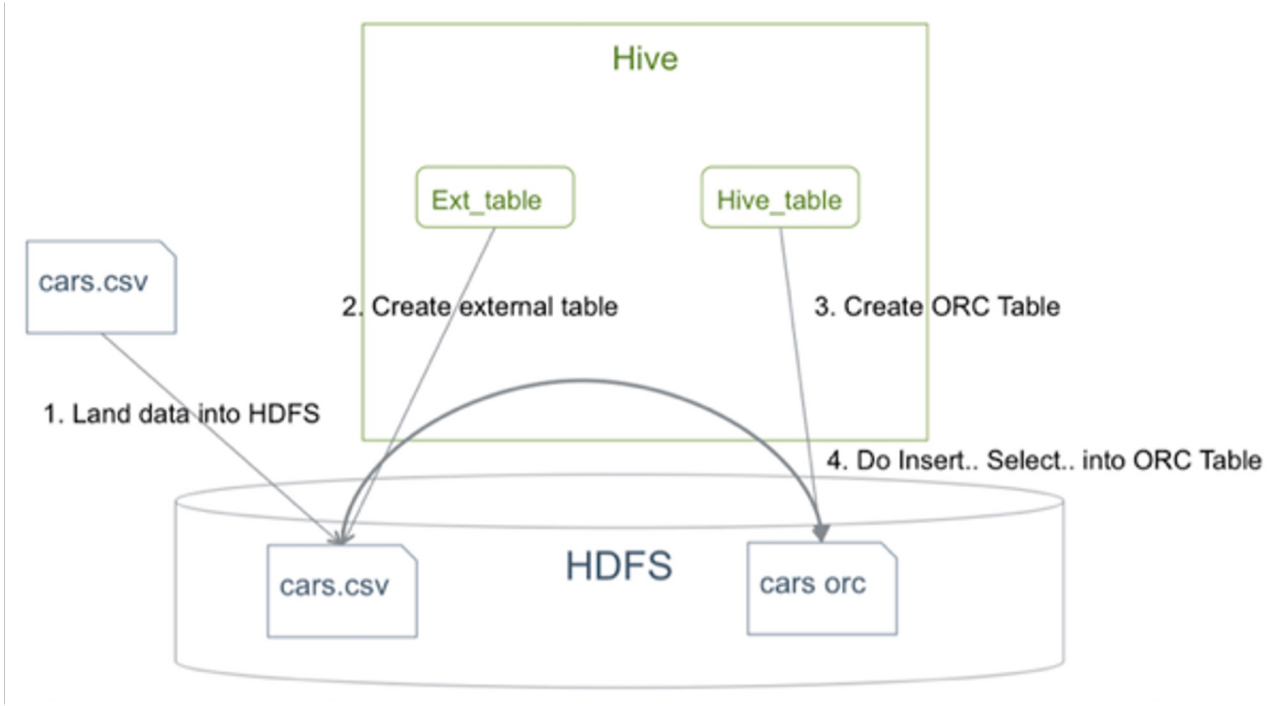
DML: INSERT

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set
  hive.exec.max.dynamic.partitions.pernode=300;
set hive.exec.max.dynamic.partitions=9000;
set hive.exec.max.created.files=18000;
```

```
INSERT [OVERWRITE|INTO] TABLE
  db.requests_logs_parquet
PARTITION (year, month, day)
SELECT
timestamp,
event_type,
request_ip,
request_port
year,
printf("%02d",month),
printf("%02d",day)
FROM staging.requests_logs;
```

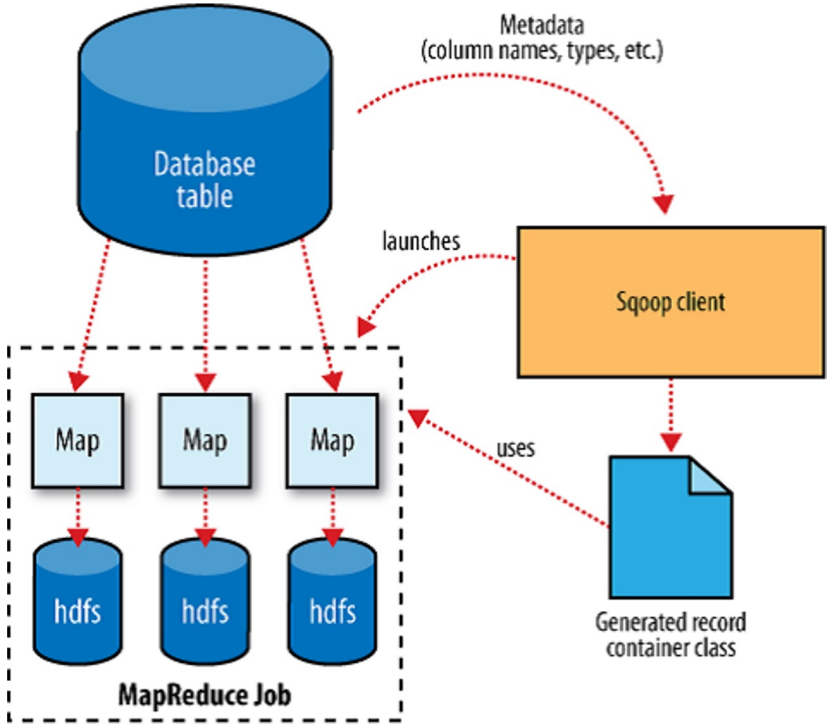
ITBA

DML: LOAD DATA

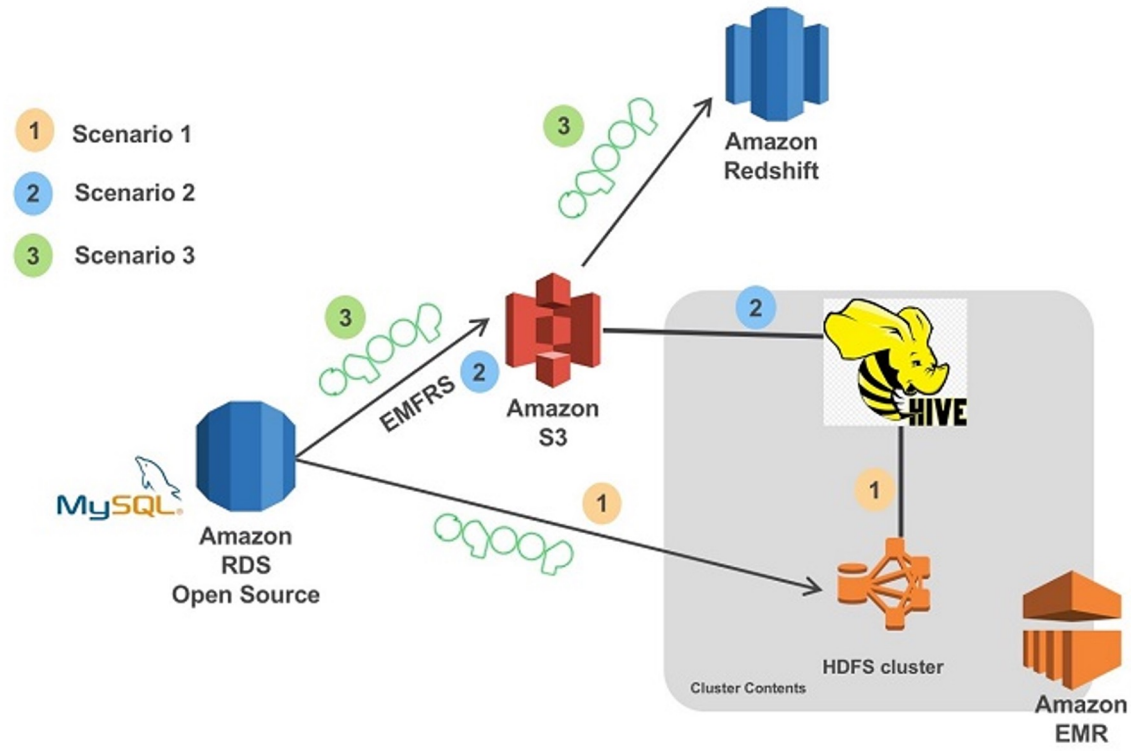


Apache Sqoop

Sqoop is a command line tool designed to transfer data between Hadoop and relational databases.



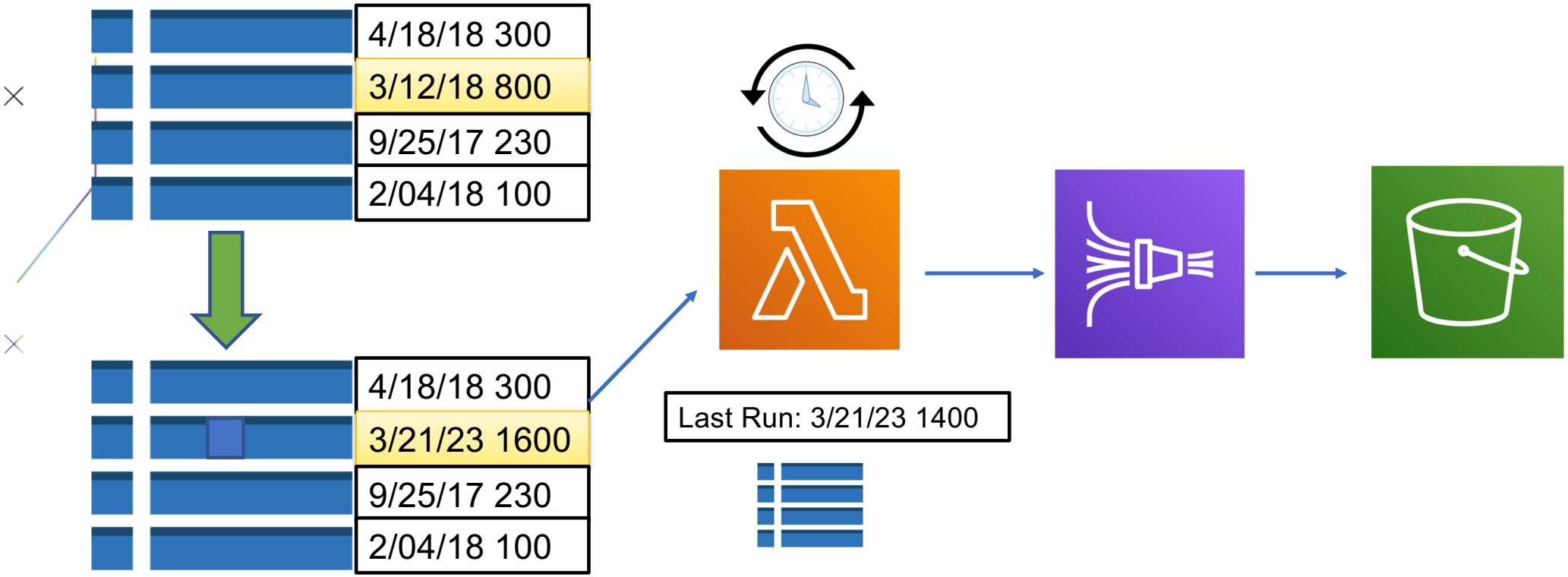
ITBA



https://aws.amazon.com/blogs/big-data/migrate-rdbms-or-on-premise-data-to-emr-hive-s3-and-amazon-redshift-using-emr-sqoop/

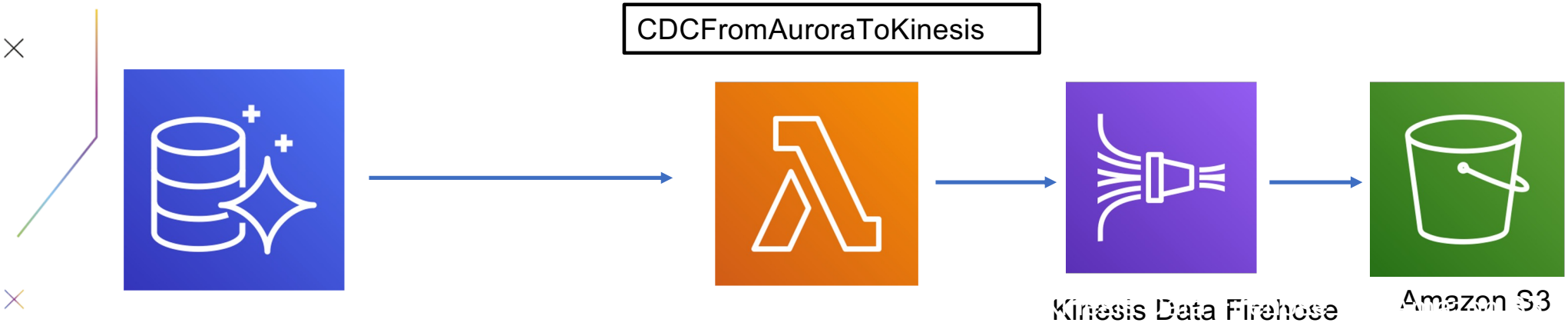
ITBA

Change Data Capture (CDC)– Timestamp using Lambda



ITBA

Change Data Capture – Triggers v2




```
CREATE PROCEDURE CDC_TO_FIREHOSE […]
      CALL mysql.lambda_async('arn:aws:lambda:us-east-
      1:XXXXXXXXXXXX:function:
CREATE TRIGGER TR_Sales_CDC AFTER INSERT ON Sales […]
      CALL CDC_TO_FIREHOSE
```

CDCFromAuroraToKinesis

Amazon Kinesis Data Firehose

- Zero administration and seamless elasticity
- Direct-to-data store integration
- Serverless continuous data transformations
- Near real-time
- Data format conversion to Parquet/ ORC

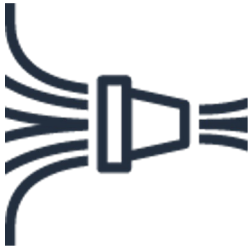
ITBA

aws

Amazon Kinesis – Streams vs Firehose

Kinesis Data Streams

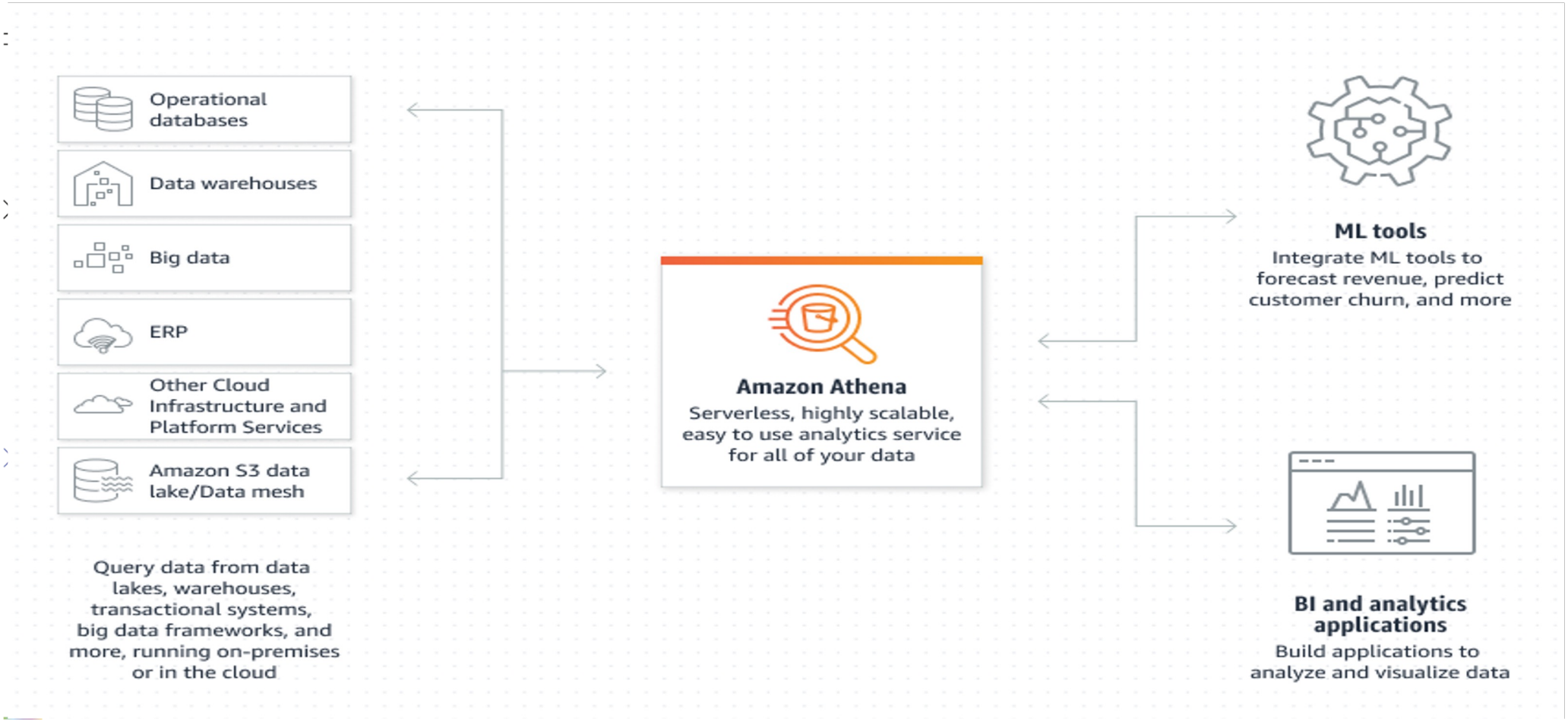
Amazon Kinesis Data Streams is for use cases that require custom processing, per incoming record, with sub-1 second processing latency, and a choice of stream processing frameworks



Amazon Kinesis Data Firehose is for use cases that require zero administration, ability to use existing analytics tools based on Amazon S3, Amazon Redshift, and Amazon ES, and a data latency of 60 seconds or higher

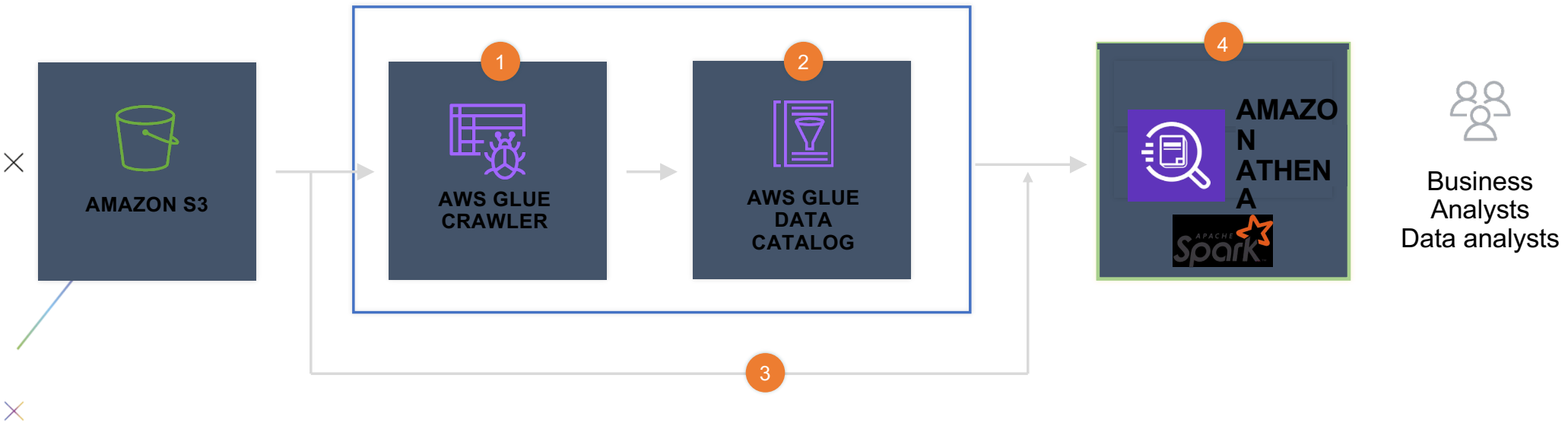
ITBA

aws


Operational databases
Data warehouses
Big data
ERP
Other Cloud Infrastructure and Platform Services
Amazon S3 data lake/Data mesh
Query data from data lakes, warehouses, transactional systems, big data frameworks, and more, running on-premises or in the cloud
Amazon Athena
Serverless, highly scalable, easy to use analytics service for all of your data
ML tools
Integrate ML tools to forecast revenue, predict customer churn, and more
BI and analytics applications
Build applications to analyze and visualize data

ITBA

Amazon Athena

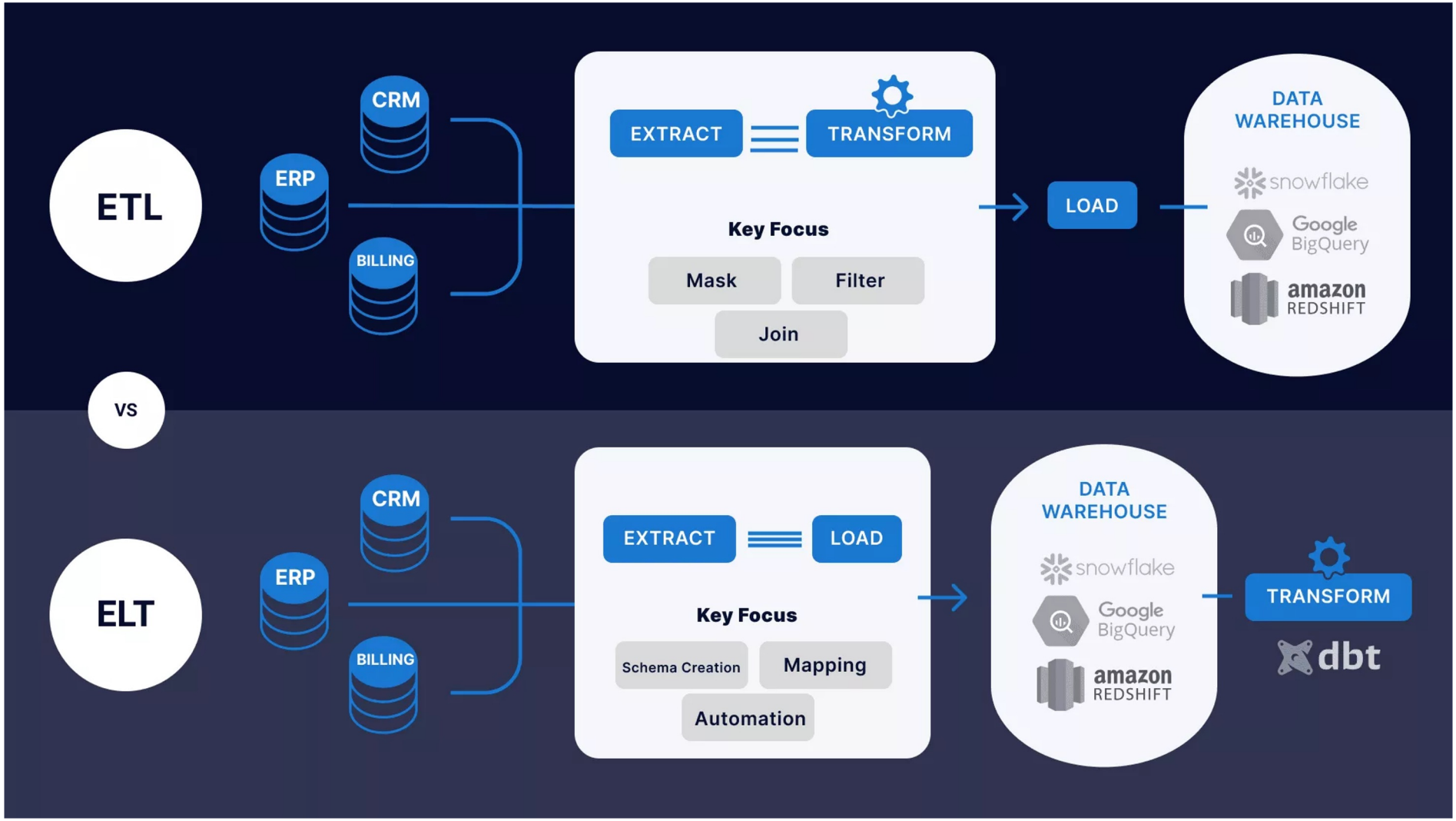


1. Crawlers scan your data sets and populate the data catalog
2. The data catalog serves as a central metadata repository
3. Alternately, directly query the data in S3
4. Query and visualize your data using notebooks in Amazon Athena for analytics

ITBA

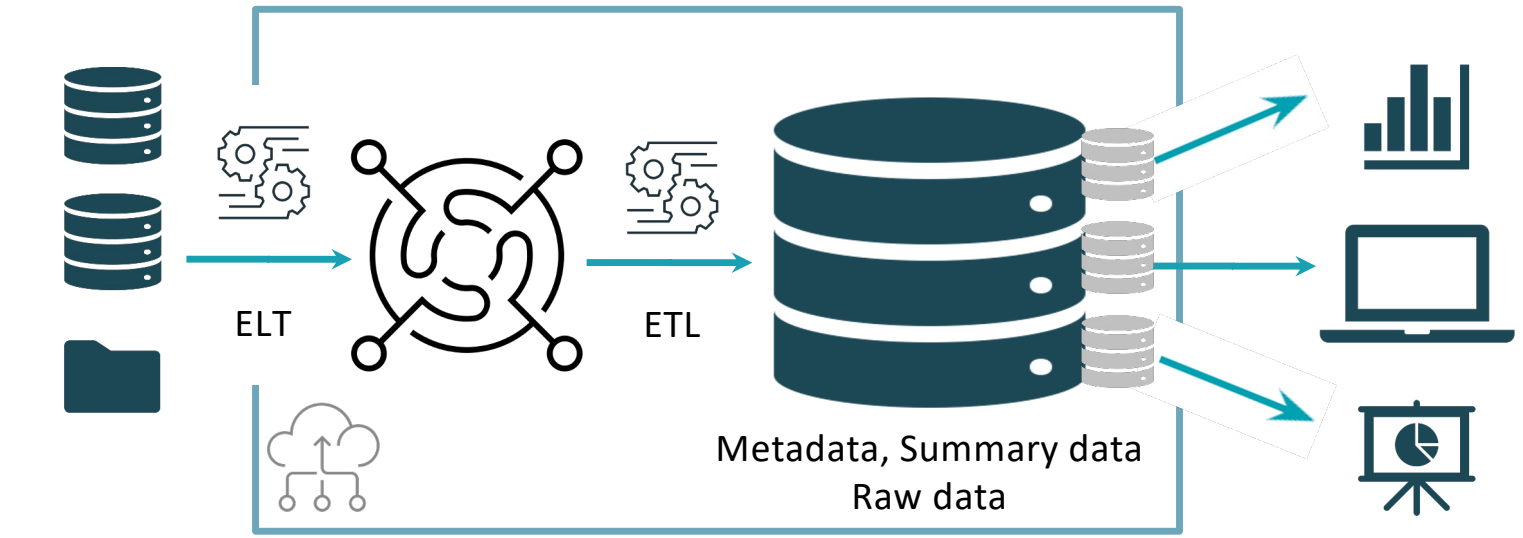
03 | Arquitecturas delta y data quality

ETL vs ELT in Modern Data Warehouse



ITBA

Cloud Based Data Lake + Warehouse



Operational ODS

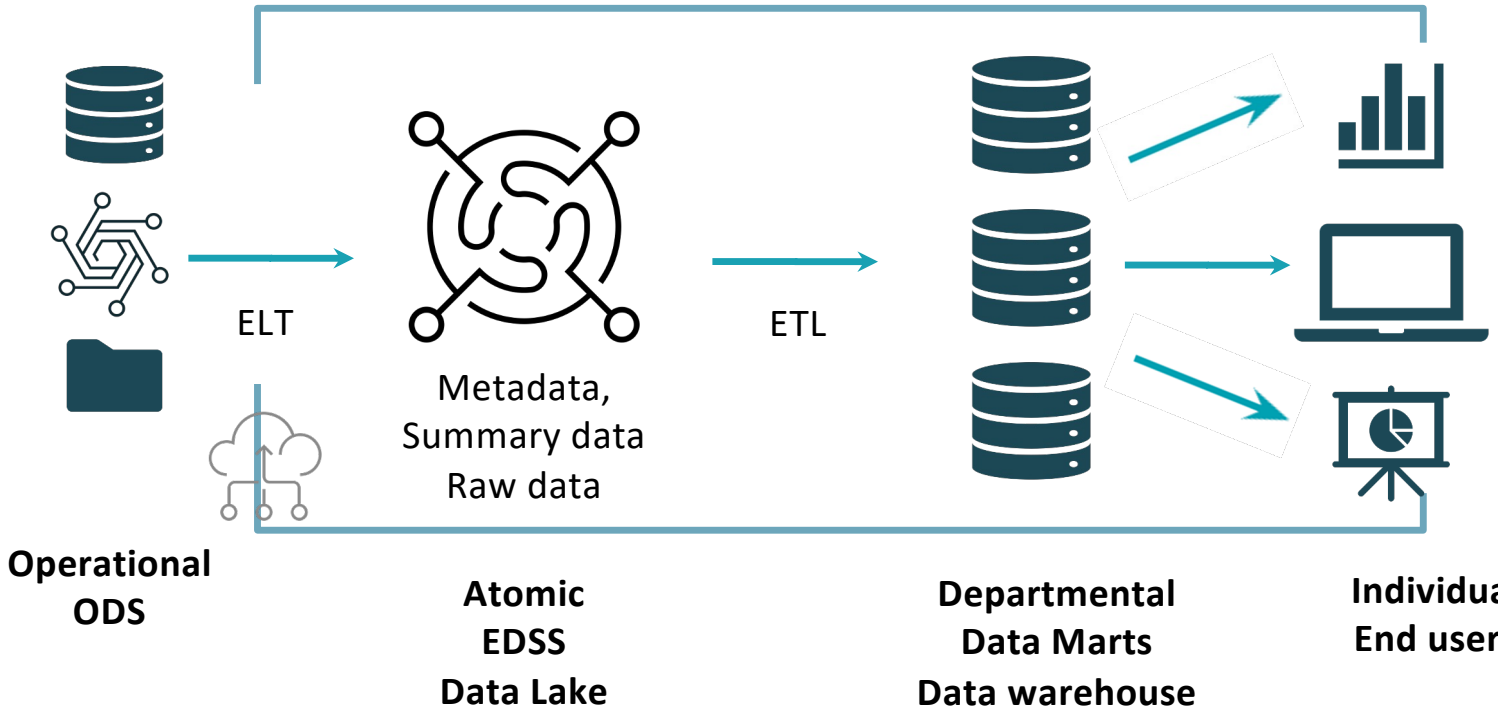
Object storage Data Lake Raw layer

Atomic EDSS

Departmental Data Marts

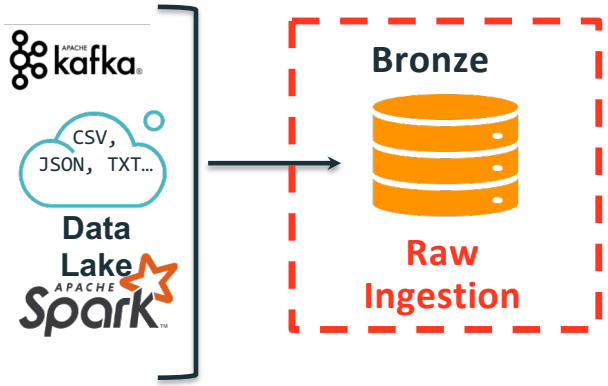
Individual End users

Delta Lake


ELT
Metadata,
Summary data
Raw data
ETL
Operational
ODS
Atomic
EDSS
Data Lake
Departmental
Data Marts
Data warehouse
Individual
End users

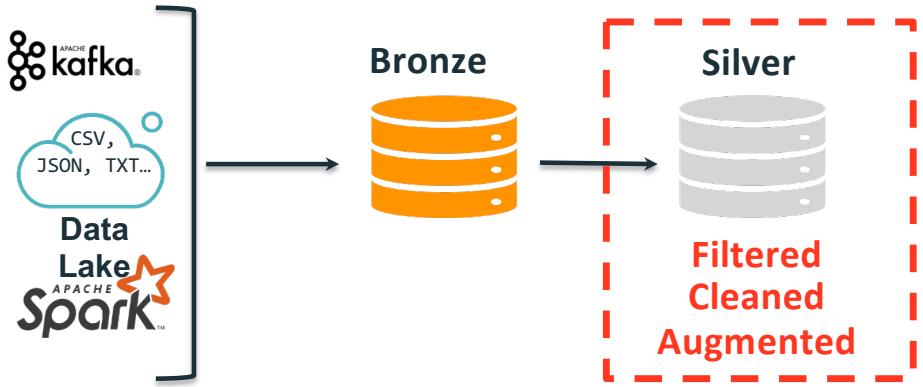
Delta Architecture Pattern

Delta Architecture: Bronze



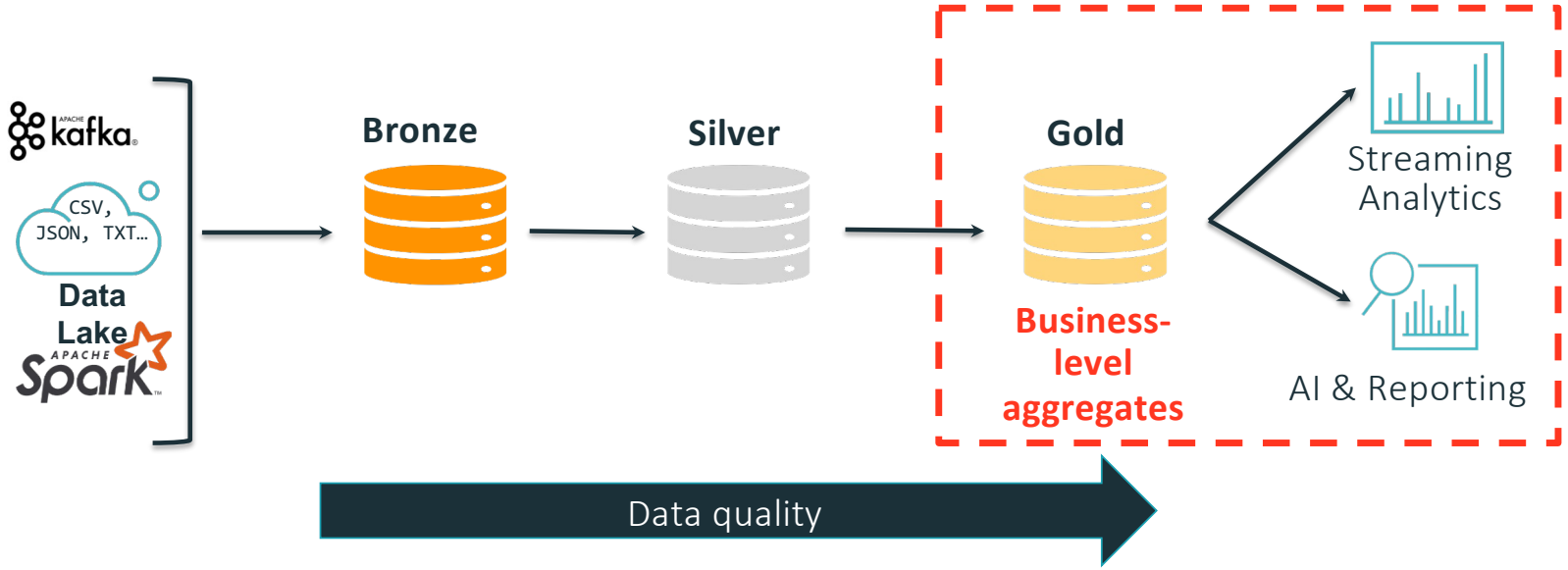
ITBA

Delta Architecture: Gold

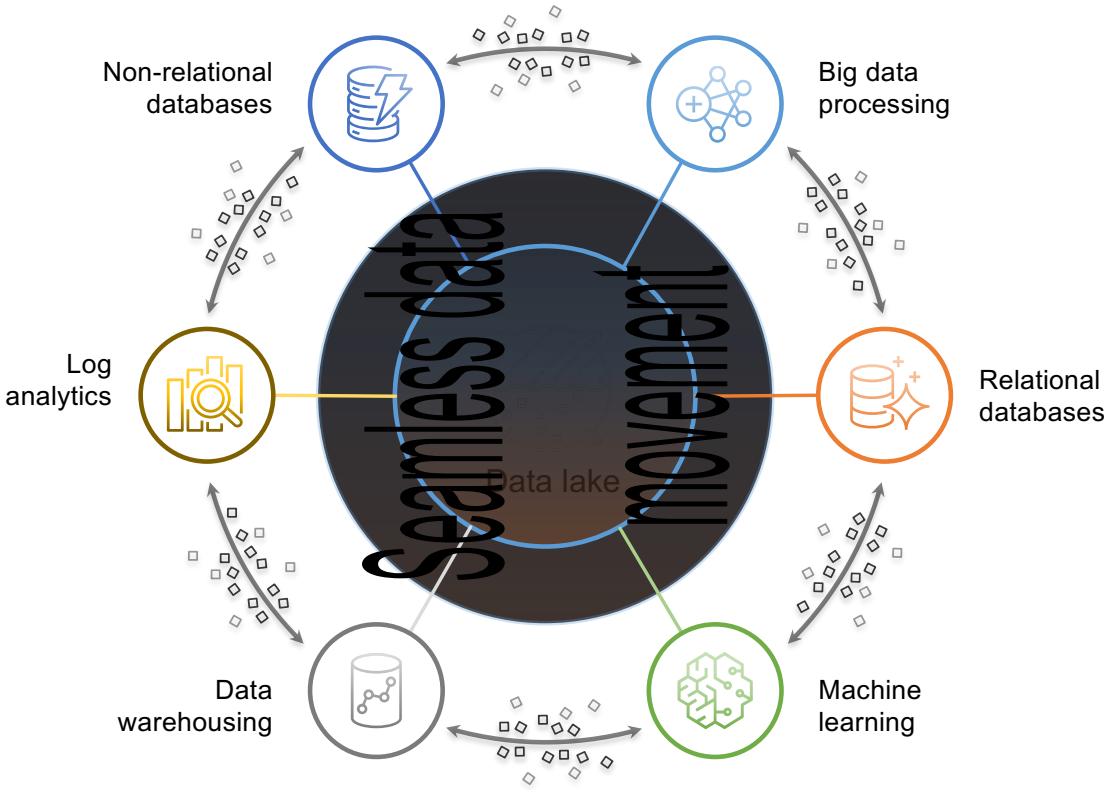


ITBA

Delta Architecture Pattern



Seamless Data Movement



04 | Práctica

Práctica

Lab 5: https://catalog.us-east-1.prod.workshops.aws/join?access-code=19e4-0f5a1f-a7
Access Code: 19e4-0f5a1f-a7

Pasos a Ejecutar:
Athena Basics Workshop
Short URL: https://tinyurl.com/596wahue

Long URL: https://catalog.us-east-1.prod.workshops.aws/work
49b5-8387-cb01d238bb78/en-US/30-basics

ITBA Escuela de INNOVACIÓN

EVOLUCIÓN CONTINUA

(+54 11) 3754-4848
(+54 11) 3754-4843

educacionejecutiva@itba.edu.ar
www.itba.edu.ar

/itbauniversidad
@itba