Apache Spark & Apache Airflow

Data Engineering Use-Cases

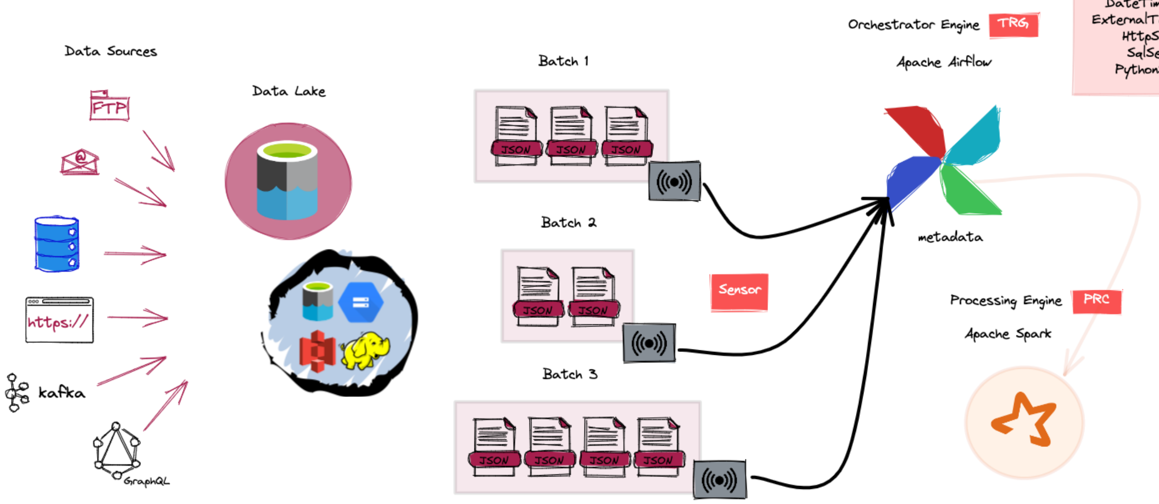
PT1

Pattern: File Sensor

Waiting for a new file to arrive, most often data landing in a data lake folder in a general format to be consumed by the spark engine to be processed.

Sensors

S3KeySensor
DateTimeSensor
ExternalTaskSensor
HttpSensor
SqlSensor
PythonSensor



● operator that check if a condition is satisfied on a particular interval
https://www.astronomer.io/guides/what-is-a-sensor/

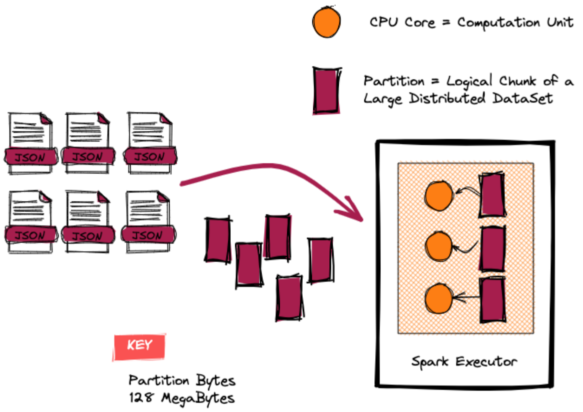
Apache Spark & Apache Airflow

Data Engineering Use-Cases

PT2

Pattern: Compact Files

After moving or copying files from the landing zone to the "processing zone" one of the best practices for spark engineers is to size the files to avoid small files problems

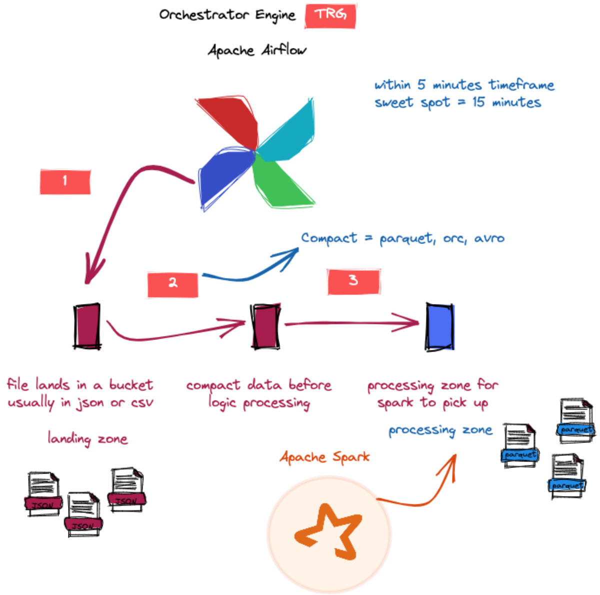


CPU Core = Computation Unit

Partition = Logical Chunk of a Large Distributed DataSet

KEY

Partition Bytes 128 MegaBytes



Orchestrator Engine TRG

Apache Airflow

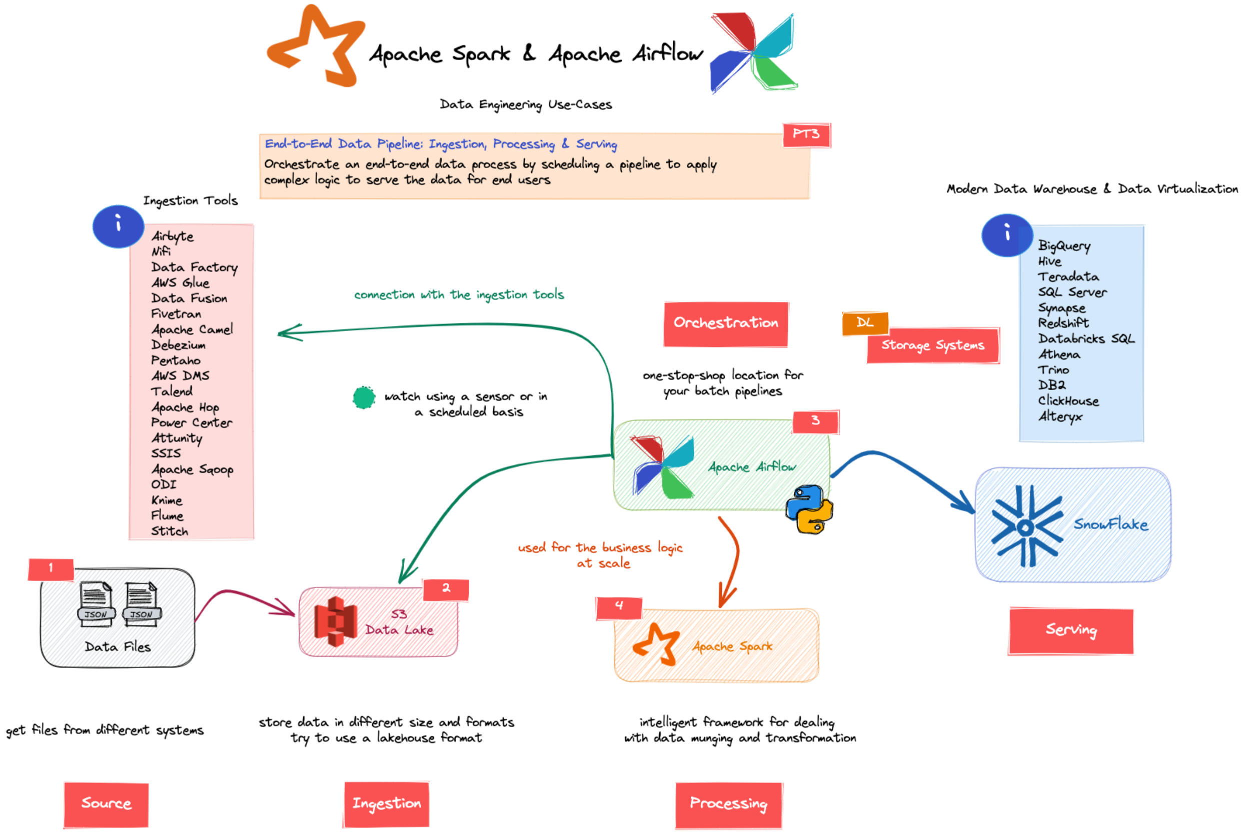
within 5 minutes timeframe sweet spot = 15 minutes

Compact = parquet, orc, avro

1 file lands in a bucket usually in json or csv

2 compact data before logic processing

3 processing zone for spark to pick up



Apache Spark & Apache Airflow

Data Engineering Use-Cases

PT3

End-to-End Data Pipeline: Ingestion, Processing & Serving

Orchestrate an end-to-end data process by scheduling a pipeline to apply complex logic to serve the data for end users

Ingestion Tools

- Airbyte
- Nifi
- Data Factory
- AWS Glue
- Data Fusion
- Fivetran
- Apache Camel
- Debezium
- Pentaho
- AWS DMS
- Talend
- Apache Hop
- Power Center
- Attunity
- SSIS
- Apache Sqoop
- ODI
- Knime
- Flume
- Stitch

Modern Data Warehouse & Data Virtualization

- BigQuery
- Hive
- Teradata
- SQL Server
- Synapse
- Redshift
- Databricks SQL
- Athena
- Trino
- DB2
- ClickHouse
- Alteryx

connection with the ingestion tools
Orchestration
DL
Storage Systems
watch using a sensor or in a scheduled basis
one-stop-shop location for your batch pipelines
3
Apache Airflow
SnowFlake
used for the business logic at scale
1
JSON
JSON
Data Files
2
S3
Data Lake
4
Apache Spark
Serving

get files from different systems

store data in different size and formats try to use a lakehouse format

intelligent framework for dealing with data munging and transformation

Source

Ingestion

Processing

Apache Spark & Apache Airflow

Data Engineering Use-Cases

Benefits

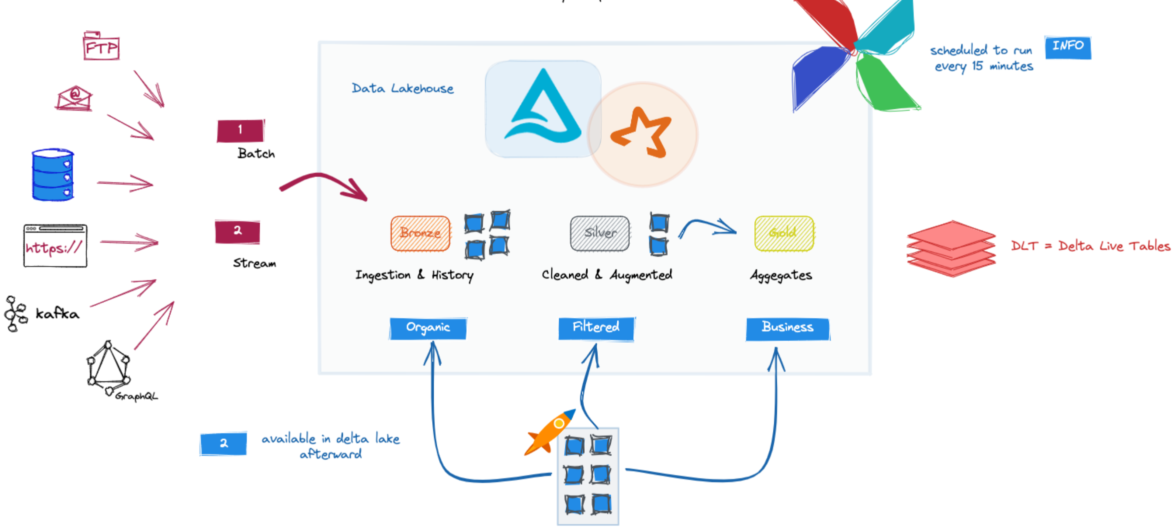
Simple Data Model
Easy to Follow
Enables Incremental
Recreate Tables
ACID Transactions
Time Travel

PT4

The Medallion Architecture

A data design pattern used to logically organize data in a lakehouse, with the the goal of incrementally and progressively improving the structure and quality

INFO Orchestrate the flow events using Apache Airflow, by leveraging the modules created by the providers available



FTP

https://

kafka

GraphQL

1 Batch

2 Stream

Data Lakehouse

Bronze — Ingestion & History

Silver — Cleaned & Augmented

Gold — Aggegates

Organic

Filtered

Business

scheduled to run every 15 minutes INFO

DLT = Delta Live Tables

2 available in delta lake afterward

NEW

Change Data Feed
delta.enableChangeDataFeed = true

https://www.databricks.com/glossary/medallion-architecture

Apache Spark & Apache Airflow

Data Engineering Use-Cases

INFO Orchestrate ingestion of different systems
to guarantee actionable insights to data teams
that needs to drive its decisions through data

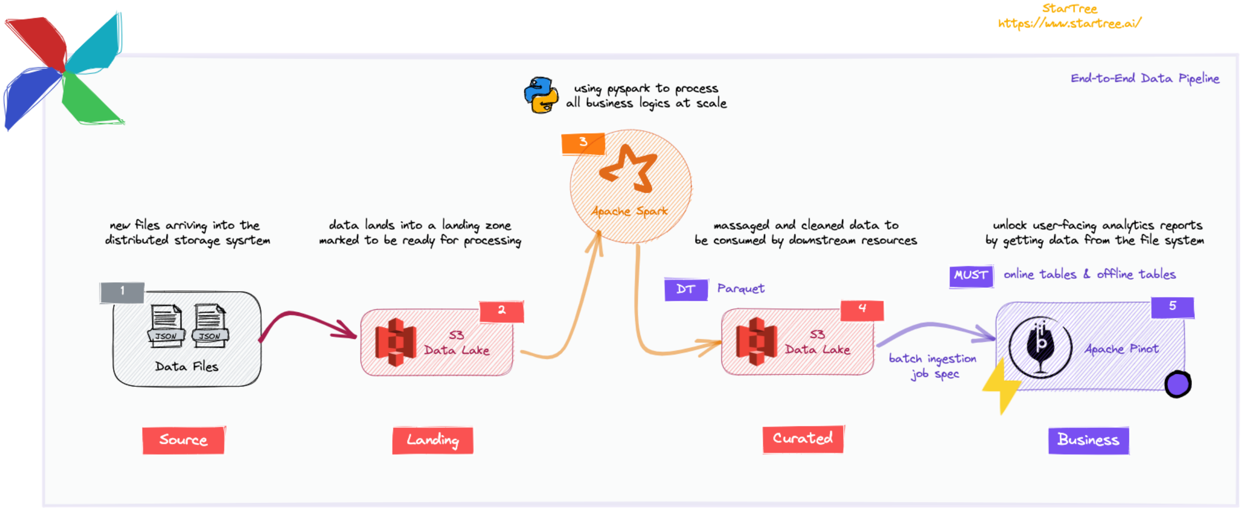
PT5

OLAP-System Pipeline Ingestion

Combine different sets of data to attend the batch side of the process.
backfilling and combining logic that needs to be applied to send data to the business users

Managed
Apache Pinot Service

StarTree
https://www.startree.ai/



End-to-End Data Pipeline

using pyspark to process
all business logics at scale

new files arriving into the
distributed storage sysrtem

data lands into a landing zone
marked to be ready for processing

massaged and cleaned data to
be consumed by downstream resources

unlock user-facing analytics reports
by getting data from the file system

KEY https://docs.pinot.apache.org/basics/data-import/pinot-file-system

Apache Spark & Apache Airflow

Provider Patterns using Astronomer Registry

Astronomer

Multi-Cloud Design

KEY

Modern Data Orchestration

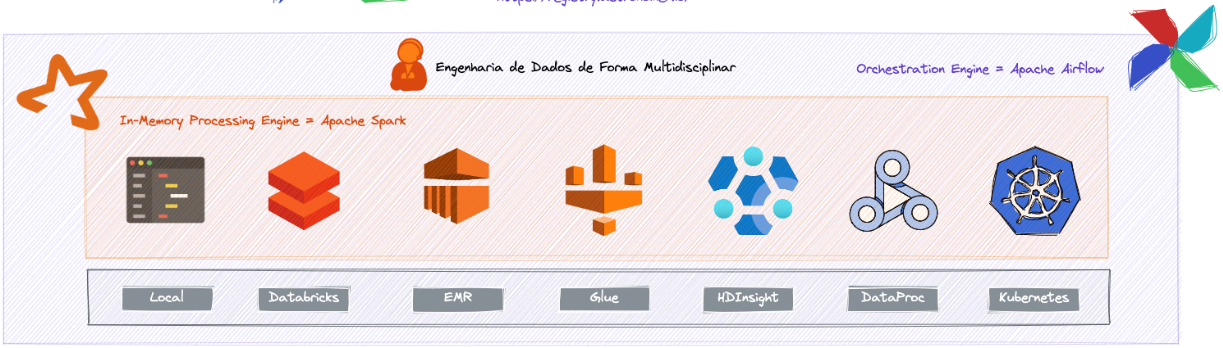
Building data flows as code using a unified platform that combines: integration, data science, operational analytics & workflow orchestration in one single spot that scales at your need

INFO

Astronomer Registry

Building blocks for your Apache Airflow Data Pipelines.
Set of Providers & Modules to allow Data Engineers to build scalable pipelines for ease.

https://registry.astronomer.io/


Engenharia de Dados de Forma Multidisciplinar
Orchestration Engine = Apache Airflow
In-Memory Processing Engine = Apache Spark
Local
Databricks
EMR
Glue
HDInsight
DataProc
Kubernetes

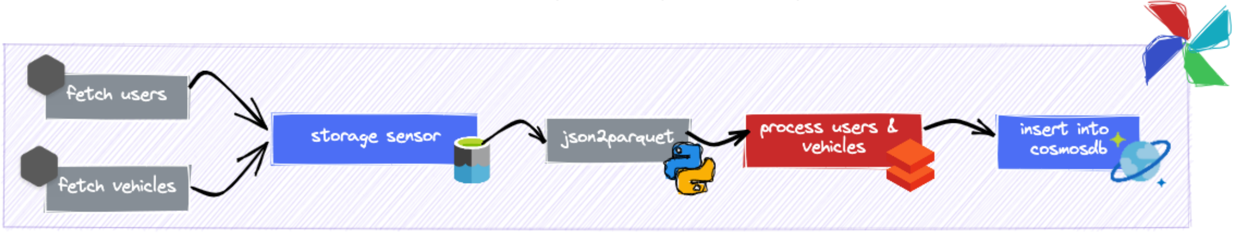
🟧 Apache Spark 🟪 Apache Airflow ⬜ Deployment Options

AKS

INFO

Data Pipeline using Apache Spark & Apache Airflow

Apache Spark = Data Harmonization
Apache Airflow = Orchestration Engine
Microsoft Azure = Backbone Infrastructure


fetch users
fetch vehicles
storage sensor
json2parquet
process users & vehicles
insert into cosmosdb