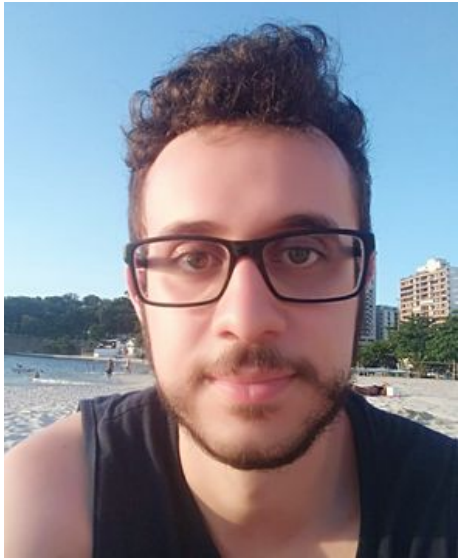
Luiz Vieira

Data Engineering Challenge

About me



Who Am I

- 26 years old, I live in São Paulo.

- I work in the data world for 5-6 years.

- Programmer since 16 years old

- I've worked with web systems development, using PHP, ASP.NET and other things that were hype around 2012.

- I graduated in economics by choice.

- Curiosity: the weirdest language I used was G-Code

Architecture proposal

Applications and services to make delivery transparent with couriers positions ... and much more.

Layers and pipeline

Goals:

Enable stream and Batch interfaces for ingestion in the Data Lake.
Configure the interface flexibity for different sources through APIs / SDKs.

Variety

Goals:

Store, transform and historicize data (delta). This step must be based on robust compression technologies, distributed computing and low latency.

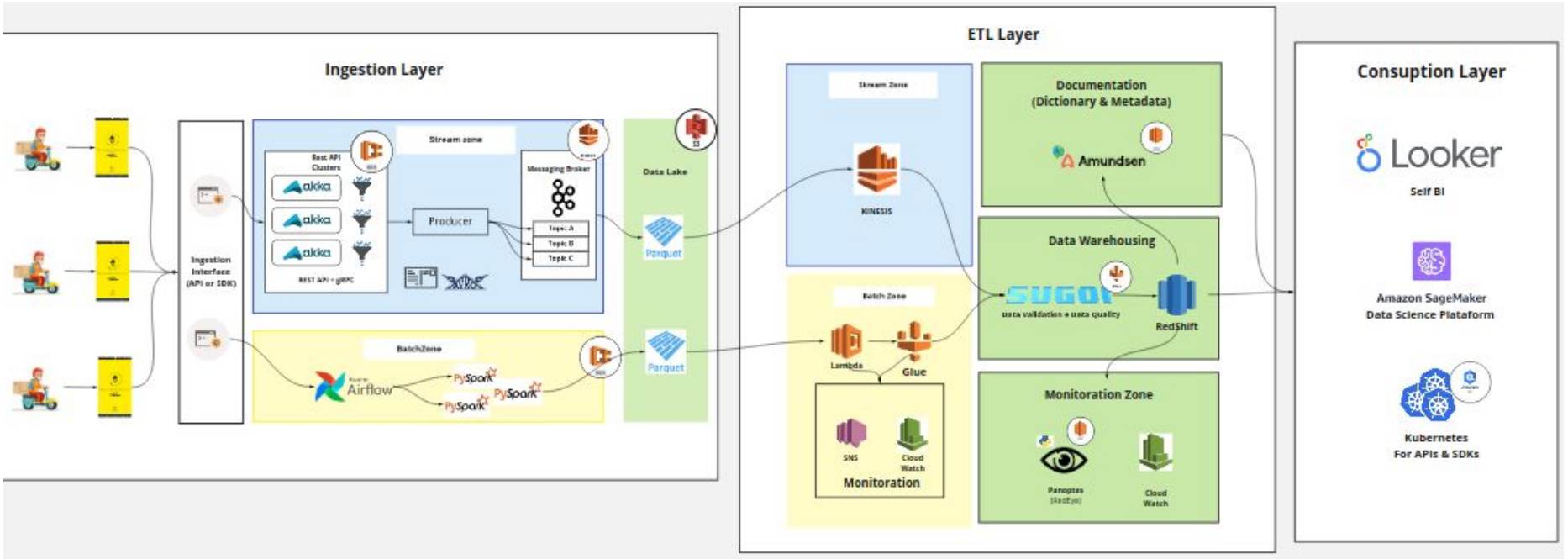
Volume

Goals:

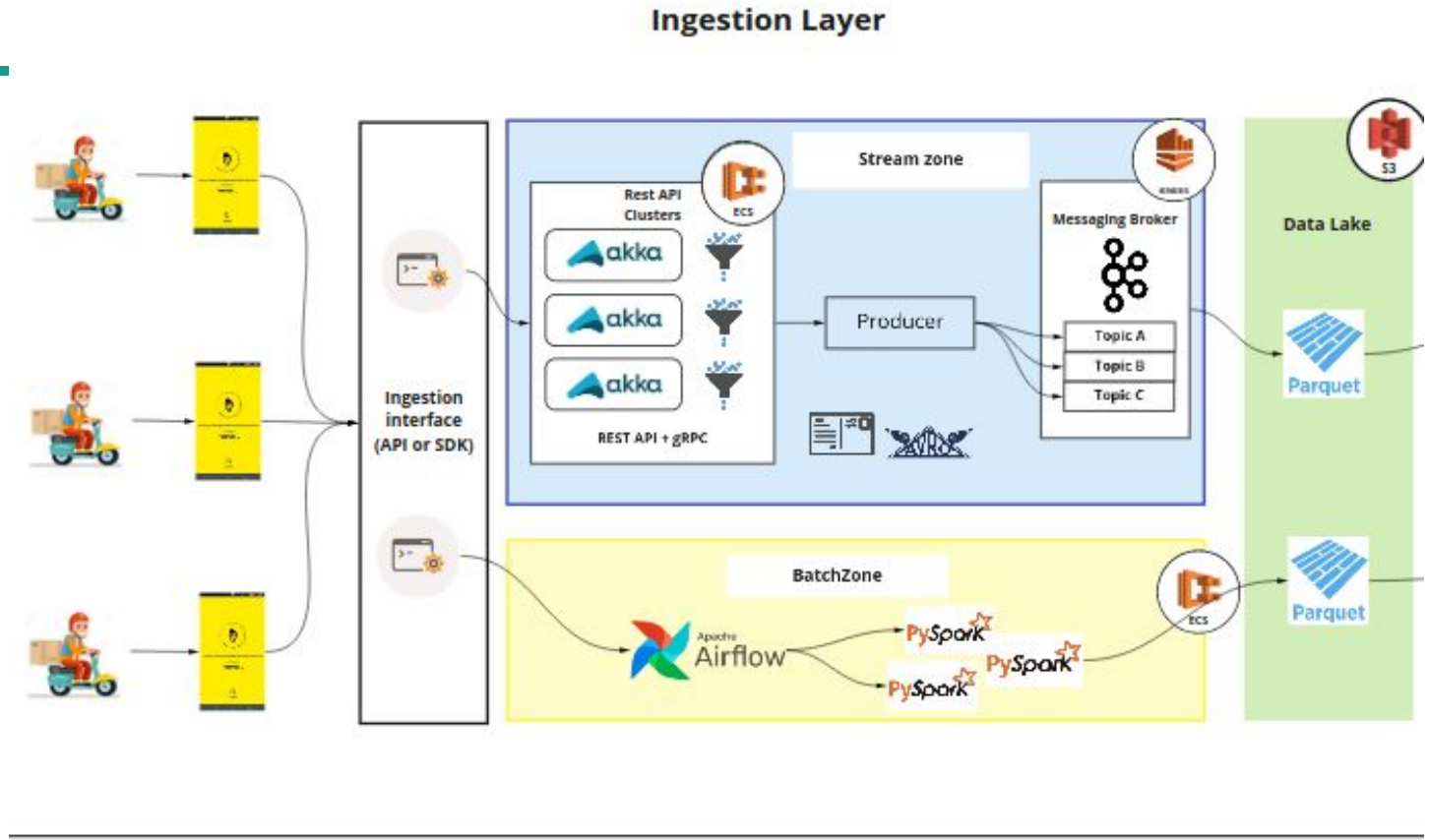
Make data available in different formats for different business needs. This step must be flexible to meet changes in the market.

Velocity

Overview


Ingestion Layer
Ingestion Interface (API or SDK)
Stream zone
Rest API Clusters
akka
akka
akka
Producer
Messaging Broker
Topic A
Topic B
Topic C
BatchZone
Airflow
PySpark
PySpark
PySpark
Data Lake
Parquet
Parquet
ETL Layer
Stream Zone
KINESIS
Batch Zone
Lambda
Glue
SNS
Cloud Watch
Monitoration
Documentation (Dictionary & Metadata)
Amundsen
Data Warehousing
SUGOI
data validation e data quality
RedShift
Monitoration Zone
Panoptes
Cloud Watch
Consuption Layer
Looker
Self BI
Amazon SageMaker
Data Science Plataform
Kubernetes
For APIs & SDKs

Ingestion


Ingestion Layer
Ingestion interface (API or SDK)
Stream zone
Rest API Clusters
ECS
akka
akka
akka
REST API + gRPC
Producer
Messaging Broker
Topic A
Topic B
Topic C
S3
Data Lake
Parquet
BatchZone
Apache Airflow
PySpark
PySpark
PySpark
ECS
Parquet

Ingestion Layer - Technologies Details

Akka provides a toolkit for the actor model compatible with kafka messaging. Has functions for cluster orchestration for APIs, providing **fault tolerance** and low latency, and a framework for the development of ingestion interfaces, providing flexibility and short time-to-market.

Avro is a file serialization algorithm, highly recommended for use with akka and kafka. can be used to compress binary or text files with high speed for stream pipelines.

kinesis is, among other things, an aws messaging service based on **kafka**. It has the advantage of being integrated with several other aws services and being quick to implement.

ECS is the aws service for container orchestration, highly recommended for application clustering. It could be to run Akka and Spark. Its advantage over EMR is the lower cost.

Ingestion Layer - Technologies Details

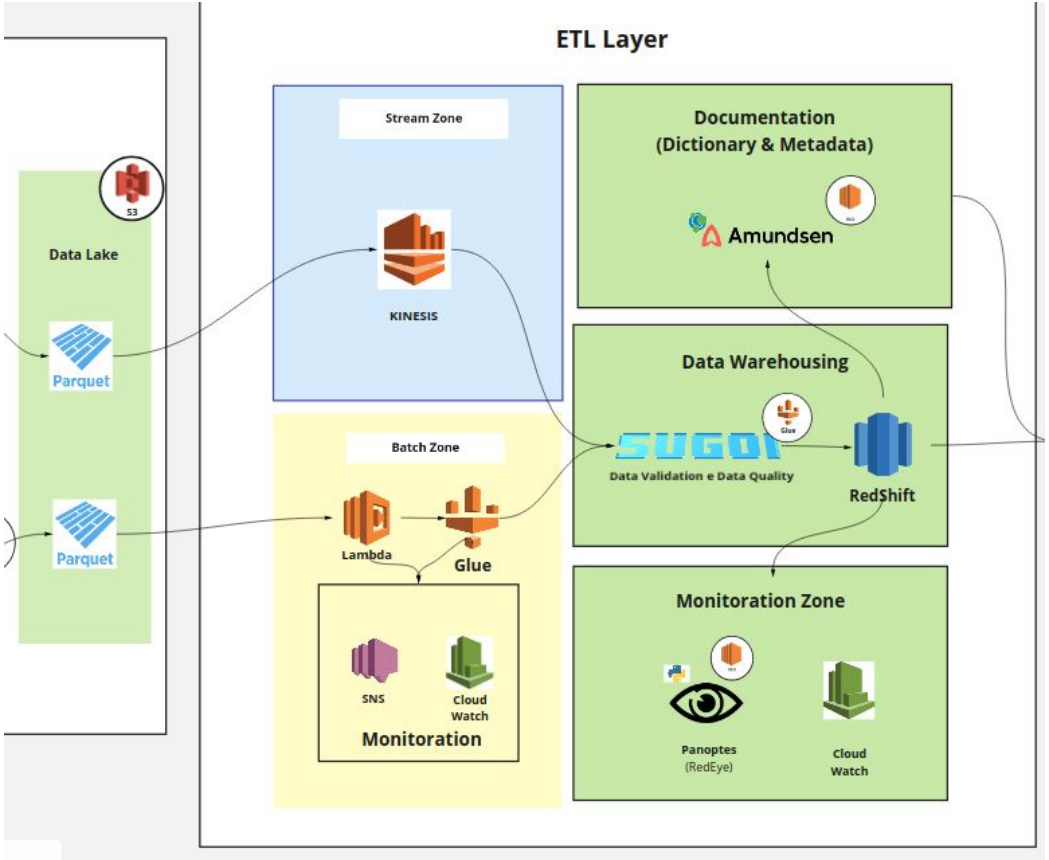
Airflow is the apache application for orchestrating complex and scalable workflows, can be written with python and has a graphical interface. It can be replaced by Luigi, which also runs in python;

S3 is the aws system for object storage, replacing the need of use HDFS. It is highly integrated with other aws services, providing easy deployment and low costs.

PySpark is an extension for spark that allows python to work with RDD in memory without the need for scala development. Spark custerization with ECS allows high speed and is highly recommended for working with big data;

Parquet is the compression algorithm maintained by the apache foundation, it uses column storage, providing compressions of more than 80% in csv files. It is highly used in big data environments due to its compression/decompression power and speed.

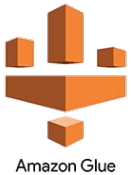
ETL


ETL Layer
Data Lake
S3
Parquet
Parquet
Stream Zone
KINESIS
Batch Zone
Lambda
Glue
SNS
Cloud Watch
Monitoration
Documentation (Dictionary & Metadata)
Amundsen
Data Warehousing
Glue
SUGOI
Data Validation e Data Quality
RedShift
Monitoration Zone
Panoptes (RedEye)
Cloud Watch

ETL Layer - Technologies Details

Lambda is a provisioning of serverless computing, which allows flexible integration between aws services. Its great differential is the low cost (none, in most cases). Amazon strongly recommends the use of lambdas to trigger other ETL services, such as Glue.

SNS iis an A2A and A2P notification system. Can be used to notify failures in the pipeline process, at any stage, for other applications and / or for users via email, API integrations and SMS



Glue is a tookit for data integration that allows you to make ETL pipelines integrated with various AWS services, such as S3, RDS, REDSHIFT, KINESIS, etc. It has cataloging engines for big data, spark environment and a built-in orchestratior. Despite its high cost, it helps to lower the costs of other services.

loudWatch is a monitoring and observation service created for DevOps engineers, developers, Site Reliability Engineers (SREs) and IT managers. Its main advantage is to have AWS services dashboard and alerts ready for use. Can be replaced by the datadog

ETL Layer - Technologies Details

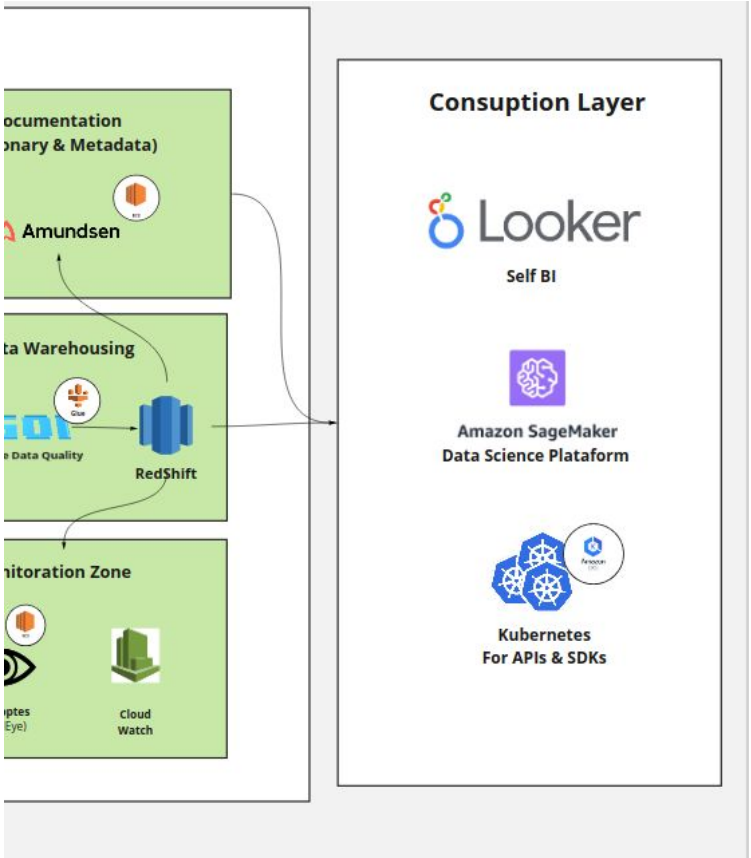
Sugoi is an open-source python library for data quality and data validation. Can be used by Pyspark running on Glue. It provides a layer that filters dirty data and provides quality reports to users.

Redshift is a fully managed, petabyte-scale data warehouse service. It can be used for streams pipelines and is highly integrated with other aws services. Despite its high cost, it provides perfectly: speed, velocity and scalability for data-driven companies.

Panoptes is a fork of the python RedEye library, which is used for monitoring hardware and software of redshift installations. Monitors the functioning of clusters with alerts (which can be replaced by Cloudwatch) in addition to monitoring queries executed, providing information such as execution time, user activity and frequent errors in queries.

Amundsen (name based on the 19th century explorer) is an open-source data discovery and metadata documentation platform, created by lyft. It works as a social network for data analysts and scientists to get to know the data contained in the data platform. It runs with Neo4j and can be integrated with hundreds of big data tools.

Consumption



ETL Layer - Technologies Details

Looker is a BI self-service tool currently maintained by google. It is extremely flexible and has a built-in modeling layer. Can be replaced by other BI tools that support stream pipeline, such as PowerBI.

Amazon SageMaker

Sagemaker is an environment for aws machine learning development. It has Jupyter notebooks for python and R and is highly integrated with services such as S3, Redshift, Athena and others.

EKS (Elastic Kubernetes Service) is a service that provides clusters for kubernete applications. Manages all nodes and automates tasks, making developers concentrate on applications. Highly recommended for applications that serve APIs, with scalabiity and fault tolerance.

This architecture are drafted in terraform code (IaC).

A document with architecture overview can be found.

Access:
https://github.com/juniorbnkr/ze-challenge

Até mais :)

Thanks :)