Data Streaming on the Cloud

with Amazon Kinesis and Spark Streaming

whoami

- Pedro - from Brazil.
- MSc in Data Science.
- Originally from Computer Engineering.
- You can find me here:

pedromb

pedromb

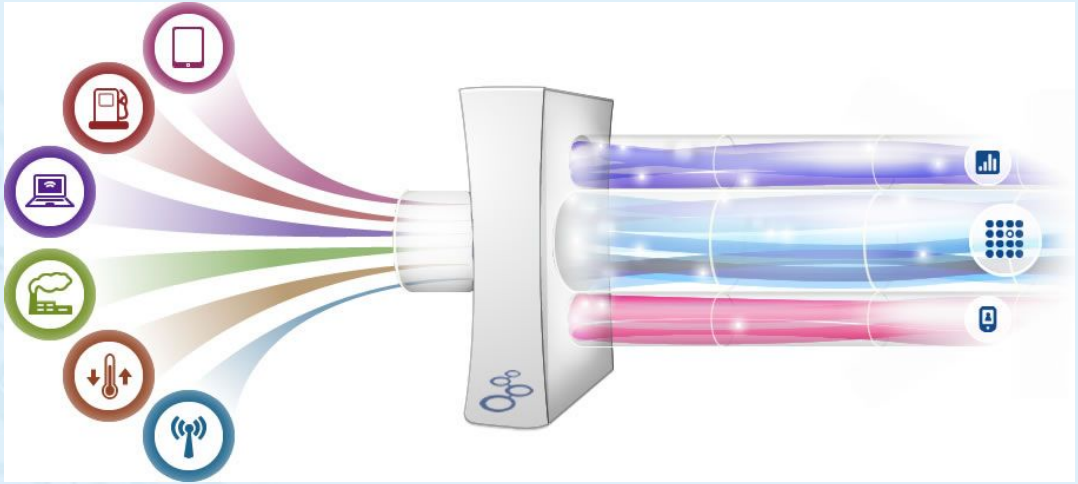
pedromagalhaesbernardo@gmail.com

Agenda

1. Data Streaming
    a. What is
    b. Benefits
    c. Challenges
2. Let’s build a data-pipeline
    a. Architecture
    b. Amazon Kinesis
    c. Apache Spark
    d. Spark Streaming
3. Example

Data Streaming - What is?

- Streaming Data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes).



Data Streaming - Benefits

- Streaming data processing is beneficial in most scenarios where **new, dynamic data is generated** on a continual basis. It applies to **most of the** industry segments and **big data use cases**.

Data Streaming - Benefits

- If we could summarise the benefits of data streaming in one sentence, this sentence would be: **real time insights**.
  - Quick reaction to operational errors (logs streaming);
  - Improved services (new opportunities - e.g.: usage of event processing algorithms);
  - Saving costs (faster response to failure);
  - Keep up with customers trends;
  - Many others

Data Streaming - Challenges

- Two layers: **storage layer, processing layer.**
  - **Storage Layer:** record ordering and strong consistency to enable fast, inexpensive, and replayable reads and writes of large streams of data.
  - **Processing Layer:** consumes data from the storage layer, runs computations on that data, and then notifies the storage layer to delete data that is no longer needed.
- You also have to plan for **scalability**, **data durability**, and **fault tolerance** in both the storage and processing layers.

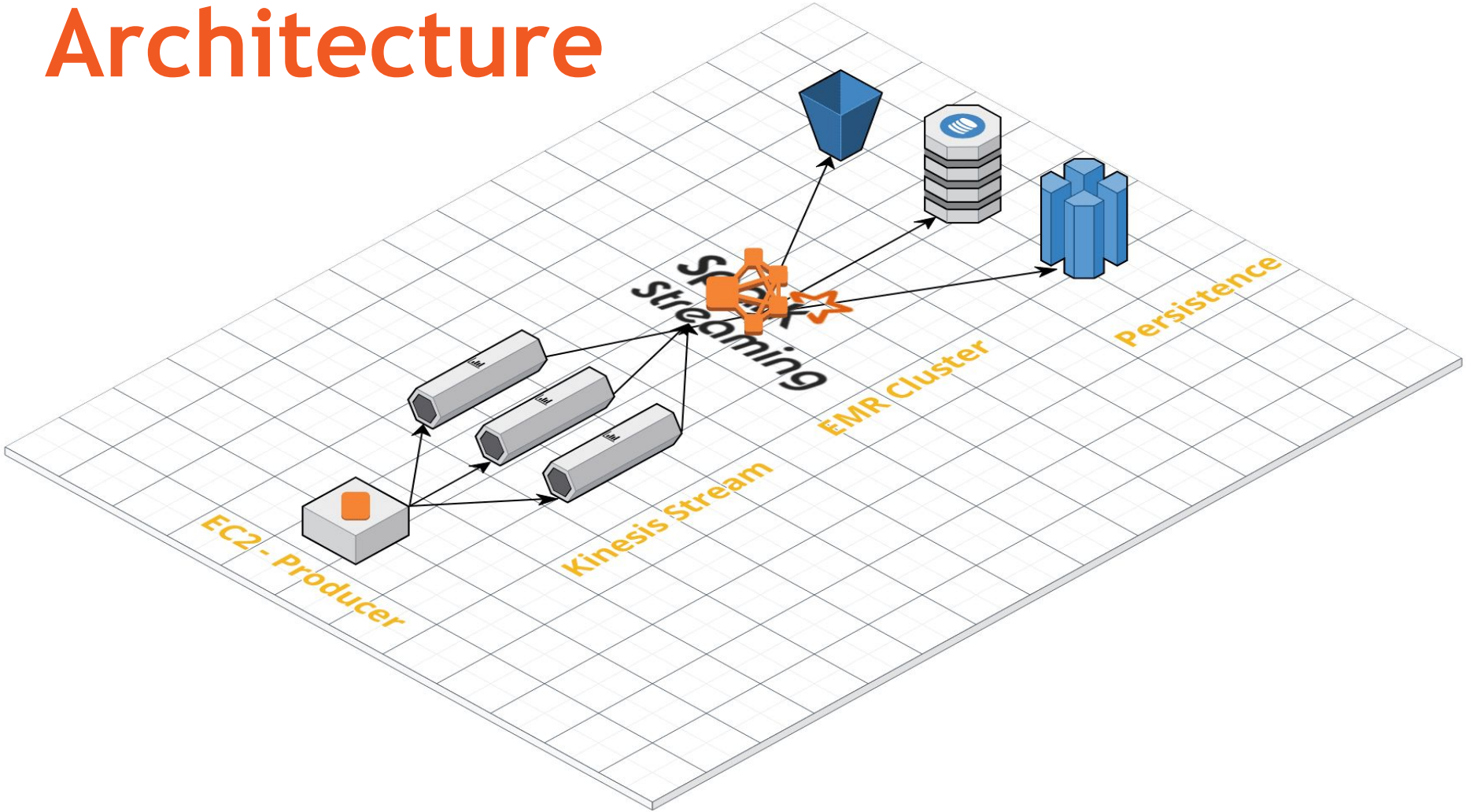
Data Streaming - Challenges

- Change on organization work flow;
- Special tools

Let’s build a Data Pipeline

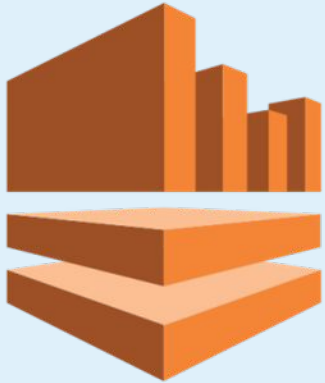
Spark Streaming

Architecture



Amazon Kinesis

Amazon Kinesis is an Amazon Web Service (AWS) for processing big data in real time.

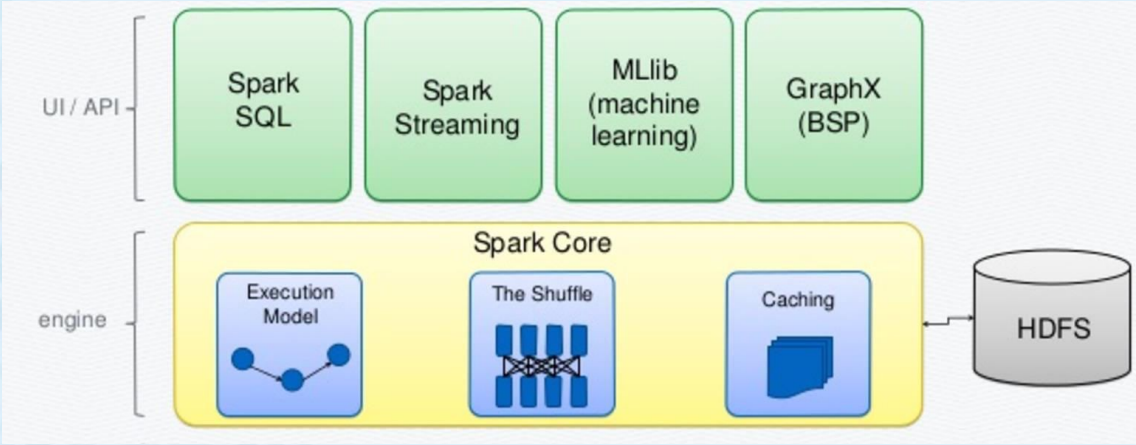


Benefits (by Amazon):

- Real-time; Secure; Easy to Use; Reliable
- Parallel Processing; Elastic; Low Cost

Apache Spark

Apache Spark is a fast, in-memory data processing engine which allows data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets.



Spark Streaming

Spark Streaming is an **extension** of the **core Spark API** that enables scalable, high-throughput, fault-tolerant stream processing of live data streams.

Data can be **ingested** from **many sources** like Kafka, Flume, Kinesis, or TCP sockets, and can be **processed using complex algorithms** expressed with high-level functions like **map, reduce, join and window**.

Finally, processed data can be **pushed out** to **filesystems**, **databases**, and **live dashboards**. In fact, you can apply **Spark's machine learning and graph processing** algorithms on data streams.

Spark Streaming

Example - Reddit

Reddit is a social news aggregation, web content rating, and discussion website.

Reddit is divided in subreddits - Each subreddit is a like a “forum” for a specific topic, it has its own moderators and rules.

For our example we will focus on the **/r/worldnews** subreddit.

/r/worldnews

30.9k

Michael Cohen Met With Qatari Official and Nuclear Plant Owner Two Days Before the FBI Raided His Offices

motherjones.com/politi...

Posted by u/singularfate 20 hours ago

1.5k Comments Share ...

13.3k

Trump As Trump-North Korea Talks Falter, South Korea Says 'Landmine' John Bolton to Blame commondreams.org/news/2...

Posted by u/maxwellhill 15 hours ago

1.4k Comments Share ...

4.6k

Trump fundraiser reportedly secured $1 billion in contracts from Saudi Arabia and UAE to push anti-Qatar policies with Trump. slate.com/news-a...

Posted by u/AdamCannon 12 hours ago

444 Comments Share ...

3.6k

Japan sexual harassment survey reveals 150 allegations by women in media:Dozens of women working for newspapers and TV have been sexually harassed – many repeatedly – with government officials, police officers and MPs cited as the perpetrators in about a third of the cases, according to a new survey theguardian.com/world/...

Posted by u/ManiaforBeatles 6 hours ago

285 Comments Share ...

Reddit API

Reddit has a great **(public/free)** API that allows querying the public data available on the website, but it has a limited rate **:(**

PRAW (Python Reddit API Wrapper) - A wrapper in Python for the Reddit API with “extra features”, such as streaming data!

We can use the streaming module from PRAW to get **real-time data** coming from a specific subreddit (we can get the comments and submissions for the /r/worldnews subreddit for example.)

Topic Modelling and Sentiment Analysis

- Not on the scope of this presentation.
- But...I am writing an article about my experience with Topic Modelling for this project, you can find more about it on my github (https://github.com/pedromb/pyDMM)
- For the Sentiment Analysis part I am using Google's Natural Language API, which is part of the Google Cloud Platform. More about it here.

In practice...

- You can find the code here: https://github.com/pedromb/data_streaming

Architecture