Elasticsearch for Hadoop

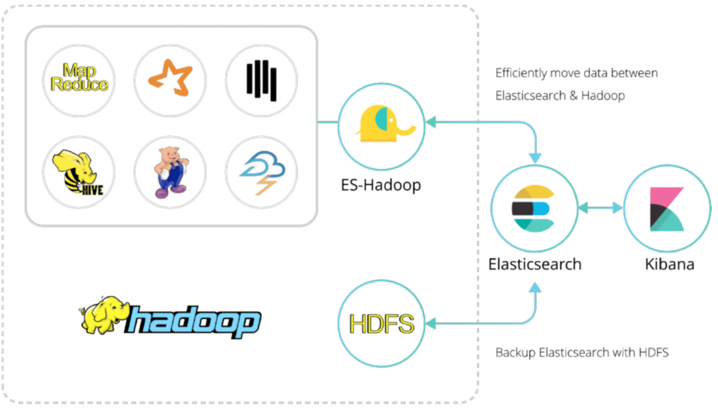
Musa Baloyi

August 20, 2018

Table of contents

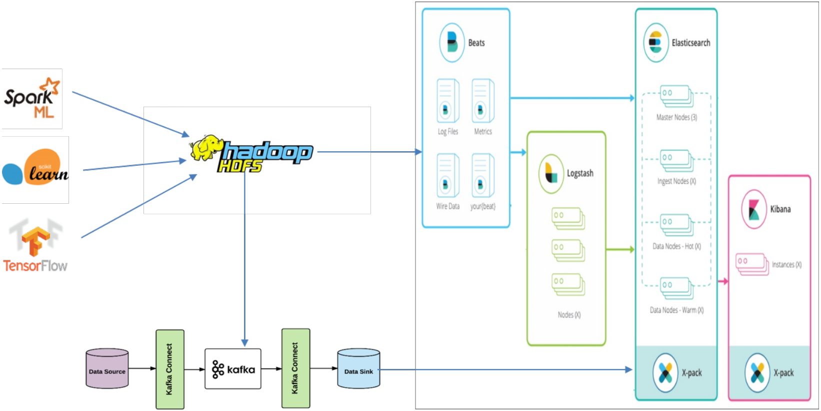
- Elasticsearch for Hadoop
- Architectural overview
- Kafka
- Gobblin
- Confluent

Elasticsearch for Hadoop



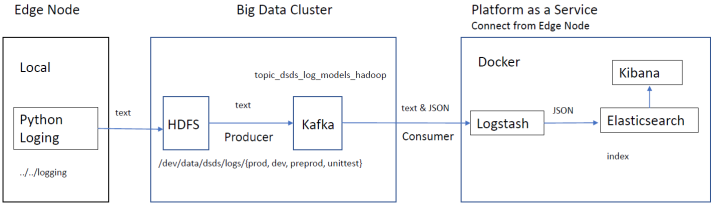
Map Reduce
HIVE
hadoop
ES-Hadoop
HDFS
Efficiently move data between Elasticsearch & Hadoop
Elasticsearch
Kibana
Backup Elasticsearch with HDFS

Architectural overview



Architectural overview (NEW)

Model Monitoring Architecture



Edge Node

Big Data Cluster

Platform as a Service

Connect from Edge Node

Local

Python Loging

../../logging

text

HDFS

text

Producer

Kafka

topic_dsds_log_models_hadoop

/dev/data/dsds/logs/{prod, dev, preprod, unittest}

text & JSON

Consumer

Docker

Logstash

JSON

Elasticsearch

index

Kibana

Kafka

- Kafka is generally used for building real-time streaming
  - data pipelines that reliably get data between systems or applications
  - applications that transform or react to the streams of data
- Kafka is run as a cluster on one or more servers that can span multiple datacenters.
- The Kafka cluster stores streams of records in categories called topics.
- Each record consists of a key, a value, and a timestamp.

Kafka installation

- Download the binary: kafka_2.12-1.0.1.tgz
- 7z x kafka_2.12-1.0.1.tgz && 7z x kafka_2.12-1.0.1.tar
- sudo mv kafka_2.12-1.0.1 /opt/Kafka

Kafka demo

- cd /opt/Kafka/ kafka_2.12-1.0.1
- sudo bin/kafka-server-start.sh config/server.properties
- bin/kafka-console-consumer.sh –bootstrap-server localhost:9092 –topic testing –from-beginning
- bin/kafka-topics.sh –create –zookeeper localhost:2181 –replication-factor 1 –partitions 1 –topic testing
- bin/kafka-topics.sh –list –zookeeper localhost:2181
- Configure Kafka producer connect-file-source.properties
- Configure Kafka consumer connect-file-sink.properties
- bin/connect-standalone.sh config/connect-standalone.properties config/connect-file-source.properties config/connect-file-sink.properties

References

1. Building Intelligent Systems: A Guide to Machine Learning Engineering. [Geoff Hulten] (Apress, 2018)
2. Trends in AI, Data Science, and Big Data. [Ben Lorica] (2017)
3. Building Evolutionary Architectures. [Rebecca Parsons; Patrick Kua; Neal Ford] (O'Reilly Media, 2017)
4. 5 things you should be monitoring. [Brian Brazil] (2018)
5. The Logstash Book. [James Turnbull] (Turnbull Press, 2013)
6. Beyond the Twelve-Factor App. [Kevin Hoffman] (O'Reilly Media, 2016)
7. Logs and real-time stream processing. [Jay Kreps] (2016)
8. I Heart Logs: Apache Kafka and Real-time Data Integration. [Jay Kreps] (2015)
9. The log: The lifeblood of your data pipeline. [Kiyoto Tamura] (2015)
10. Understanding the ELK stack. [Brian Anderson; Rafał Kuć] (2016)