Scalable Data Aggregation

MSDS697 Distributed Data Systems
Group 1 | Caleb Hamblen, Tatshini Ganesan, Rishi Mohan, Yi-Fang Tsai

Datasets

- San Francisco 311 service requests updated daily since July 2008

  (~6.8 million records)

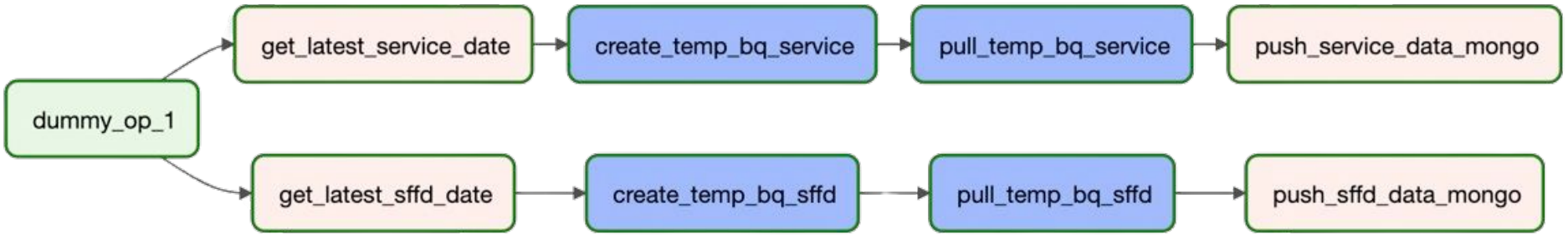
- Incidents from the San Francisco Police Department (SFPD) Crime Incident Reporting system from January 2003 until 2018

  (~2 million records)

- SF Fire Department service calls updated daily since April 2000

  (~6.5 million records)

Pipeline - Ingestion



Retrieve latest inserted date from MongoDB collection.

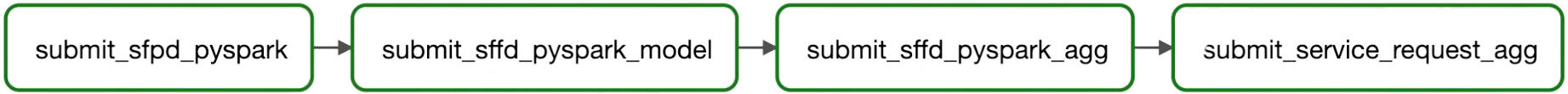
Create temporary BQ table with newer records.

Retrieve all records from temporary BQ table.

Insert new records to MongoDB collection.

Runtime:
2:45 on avg.

Pipeline - Aggregation



spark-submit job 1

`sfpd_incidents`

Various aggregations.
Writes results to
`sfpd_aggregations`.

spark-submit job 2

`sffd_service_calls`

Fits a Prophet
time-series model.
Forecasts monthly
values of a query.
Writes results to
`sfpd_aggregations`.

spark-submit job 3

`sffd_service_calls`

Various aggregations.
Writes results to
`sfpd_aggregations_2`.

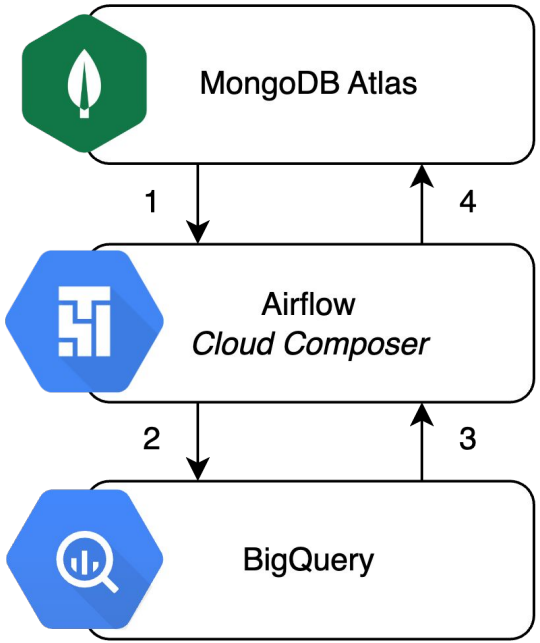
spark-submit job 4

`service_requests`

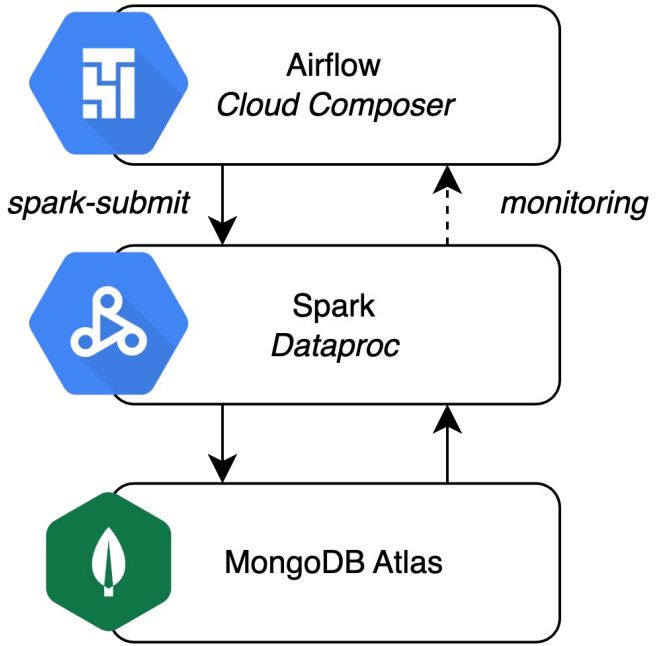
Various aggregations.
Writes results to
`service_requests_aggregations`.

Runtime:
7:13 on avg.

System Diagram



Ingestion DAG
runs daily



Simulated Modeling
runs weekly

Time Series Model

- Facebook's Prophet model forecasting the number of monthly paramedic calls resulting in a trip to the hospital

- Outperformed ARIMA model