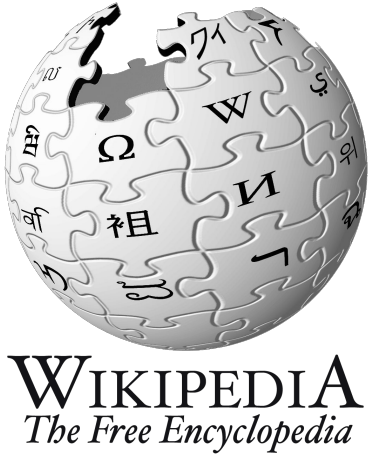
WikiStats

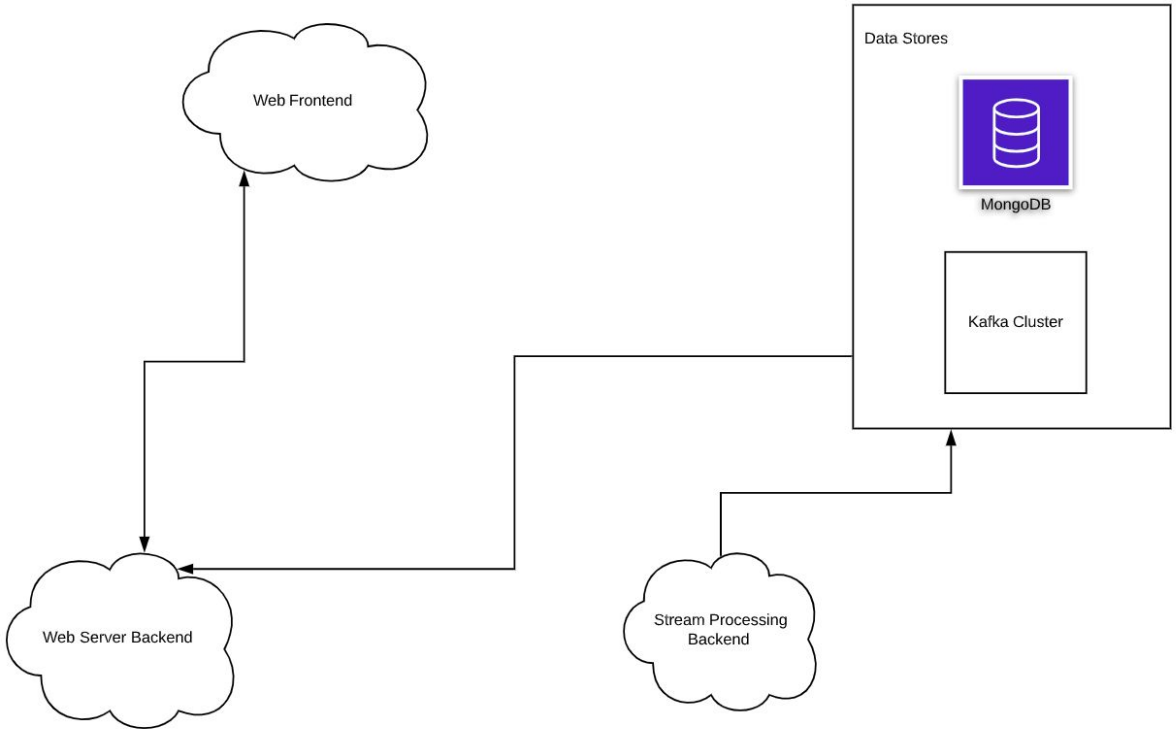
Jeswanth Yadagani (jy3012)
Sidharth Bambah (sb4283)

Motivation

- Wikipedia is a huge encyclopedia and topic database
- Useful to track growth of different wikis
- Easily understandable visual dashboard
- Provide insight for resource allocation



System Architecture



Tooling

- Stream Processing and Analysis
  - Apache Spark Streaming
  - Apache Kafka Pub/Sub Messaging
- Persistent Storage and Processed Feeds
  - MongoDB
  - Processed Kafka topic
- Web Design
  - NodeJS and Flask API
  - ReactJS and ChartJS

Raw Data

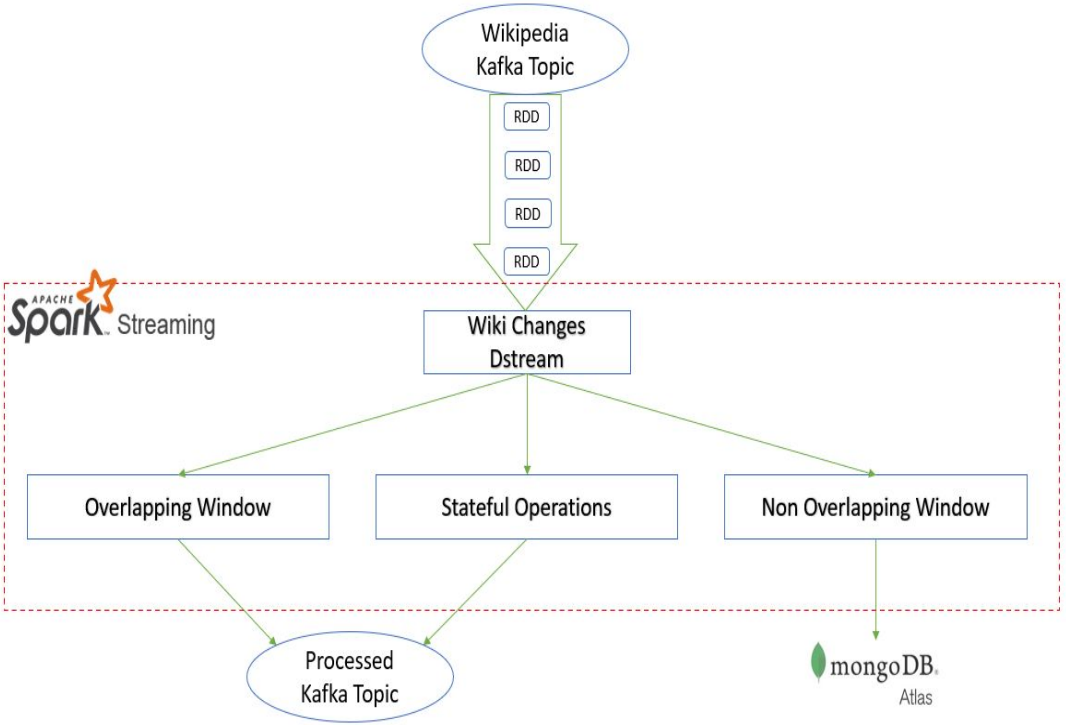

```
{
  "id":[
    {
      "topic":"eqiad.mediawiki.recentchange",
      "partition":0,
      "timestamp":1588195817001
    },
    {
      "topic":"codfw.mediawiki.recentchange",
      "partition":0,
      "offset":-1
    }
  ],
  "data":{
    "$schema":"/mediawiki/recentchange/1.0.0",
    "meta":{
      "uri":"https://en.wikipedia.org/wiki/List_of_cellists",
      "request_id":"Xqnx6QpAIHkABChLmE8AAABP",
      "id":"70f31b10-6474-4946-ab82-1a13bdca67df",
      "dt":"2020-04-29T21:30:17Z",
      "domain":"en.wikipedia.org",
      "stream":"mediawiki.recentchange",
      "topic":"eqiad.mediawiki.recentchange",
      "partition":0,
      "offset":2363500816
    },
    "id":1256260451,
    "type":"edit",
    "namespace":0,
    "title":"List of cellists",
    "comment":"",
    "timestamp":1588195817,
    "user":"142.116.134.92",
    "bot":false,
    "minor":false,
    "length":{
      "old":29959,
      "new":30038
    },
    "revision":{
      "old":953628005,
      "new":953944783
    },
    "server_url":"https://en.wikipedia.org",
    "server_name":"en.wikipedia.org",
    "server_script_path":"/w",
    "wiki":"enwiki",
    "parsedcomment":""
  }
}
```


Pre-Processed Data

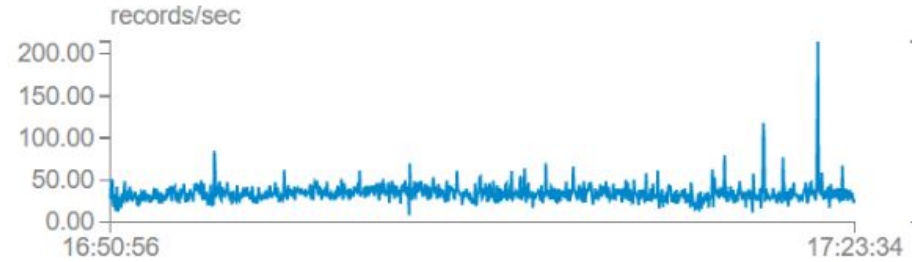

```
{
  "Class":"deletions",
  "Type":"edit",
  "Domain":"es.wikipedia.org",
  "Title":"Historia del fútbol",
  "BOT":"True",
  "User":"SeroBOT",
  "Timestamp":1588195388,
  "Comment":"Revertidos los cambios de [[Special:Contributions/177.239
  "Topic":"eqiad.mediawiki.recentchange",
  "Wiki":"eswiki"
}
```

Stream Processing

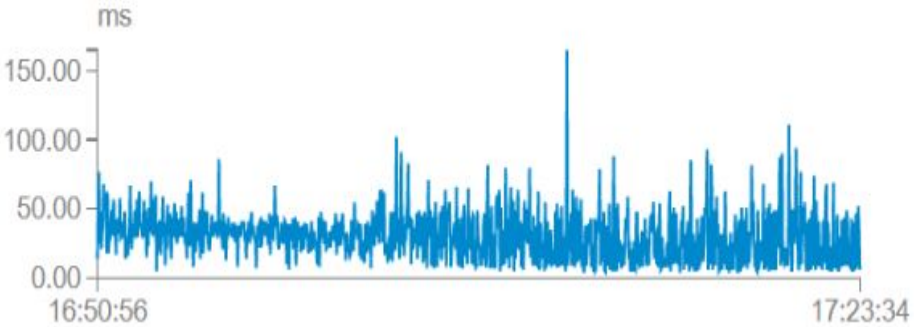
- Overlapping Window:
  - Duration: 40 sec
  - Slide: 6 sec
  - Represents trend progression
- Stateful Operations:
  - Gathers aggregate information
- Non Overlapping Window
  - Duration: 1 hr
  - Gathers information over an hour
  - Can be utilized for offline analysis



Metrics

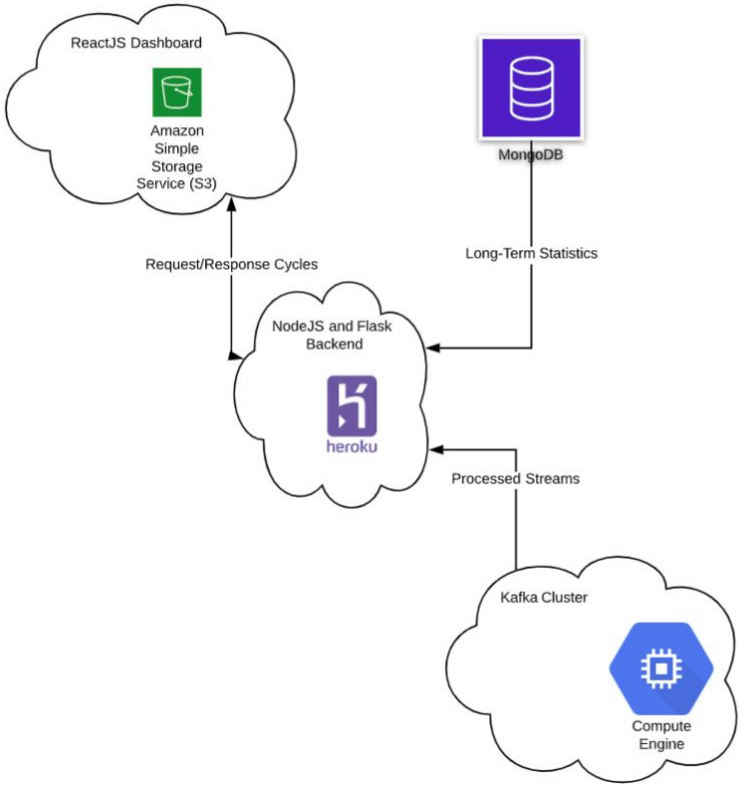


Input Rate - 34 rec/sec.
2 Sec Batch Interval



Processing Time - 50ms/batch

Web Architecture



Cloud Deployment

- Kafka Cluster: Google Compute Engine
- Stream Processing Job: Google Dataproc Cluster
- NodeJS Backend: Heroku
- Flask Backend: Heroku
- ReactJS Frontend: AWS S3

Thank You!

DEMO TIME!

Raw Data: https://stream.wikimedia.org/v2/stream/recentchange

Job in Dataproc: Stream Processor
Data from Pipeline: Start a consumer
Website Link: bit.do/wikistats