NYC Bike share analysis

[Proposal]

Aruna Kumaraswamy
Indiana University
akumaras@iu.edu

ABSTRACT

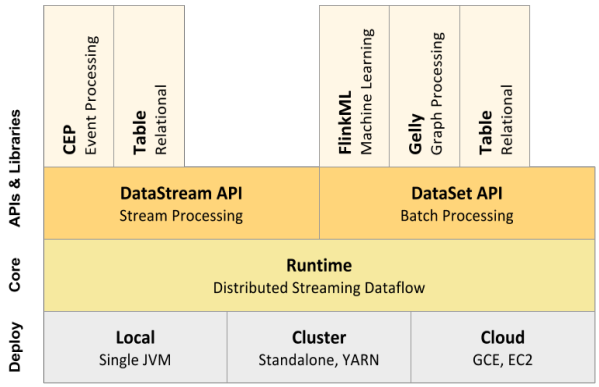
Purpose of this project is to analyze the NYC Bike share system (Citi Bike). Citi bike has published data sets for Trip history. In this project, I will be using the historical data for trend analysis. Citi Bike also publishes the availability of bikes in stations as a real time data. I will be using this streaming data for interactive visualization.

1. PROPOSAL

[**?**]Lambda Architecture is a useful framework to think about designing big data applications. Nathan Marz designed this generic architecture addressing common requirements for big data based on his experience working on distributed data processing systems at Twitter. Some of the key requirements in building this architecture include:

- Fault-tolerance against hardware failures and human errors
- Support for low latency querying as well as updates
- Linear scale-out capabilities
- Extensibility - system can accommodate newer features easily

Project will use Lambda Architecture to facilitate analysis on both real time and batch data. Spark streaming, Kaka, Hadoop and Hbase could be used to implement a lambda architecture. In this project, I would like to try an emerging technology framework for Lambda architecture - Apache Flink.



Copyright: Apache Software Foundation

Flink has both stream processing and batch environment inbuilt in the framework. It offers connectors to Kafka and Elastic Search which will be used in this project. Instead of HBase as data sink, I am using CSV data sink to explain the flow of data in this project.

1.1 Use case

Citi Bike publishes trip data for trend analysis. Trend analysis will be done in the batch mode. Result of the batch program will be used in the visualization.

Batch use cases are:

- Bike rent trends by gender
- Bike rent trends by age (generation grouping)

Citi Bike real time feed will be streamed into a Kafka topic for storing in data sink. In addition, data will be sent to an alert Kafka topic which will check the bike availability in stations against a thresold and publish alert to elastic search. Kibana will be used to visualize the alerts in elastic search.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235