Steven Black

3/23/2018

Visualizing Streaming Data  
with Kafka, Spark, and OAC

Information Management HUB Team – Reston HUB

Contact: renuka.uttarala@oracle.com

Contents

[Overview 2](#_Toc506468252)

[Putting together the Environment 2](#_Toc506468253)

[Requirements 2](#_Toc506468254)

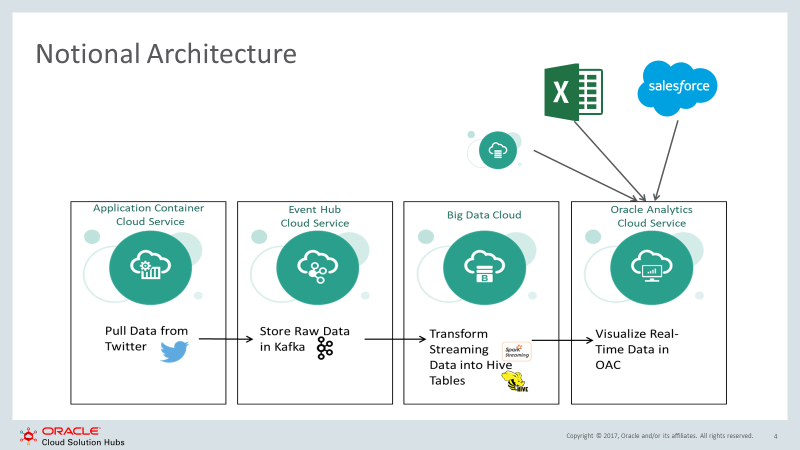
[References 4](#_Toc506468255)

# Overview

This is the technical guide to accompany the Streaming Analytics demo.

As the world digitizes data has been moving to streams. Experts have (half-jokingly) claimed “ETL is dead; Long-Live Streams”, but how can I implement an enterprise-grade solution?

Oracle offers managed services that can scale as needed. We can use ACCS to deploy a Tweet-Producer application, EHCS to store the raw streams for a set period, BDC to transform and store that data, and OAC to visualize the data as well as do additional Machine Learning.



# Putting together the Environment

This has been almost entirely streamlined with bootstrap.sh scripts. Just put the script in the object store container attached to the respective service and it will be run during creation. Alternatively, you can SSH into the service and run them yourself! Nonetheless, we will need to instantiate 5 services if this has not been done automatically.

## Getting Started

### Cloud Components

* OAC (1 OCPU) – Visualize the Data
  + DB (1 OCPU) – Necessary for OAC
* BDC (2 OCPUs) – Transform and Store the Data
* Kafka (1 OCPU) – Store Raw Data
* ACCS (1 OCPU) – Pull the Data

Total: 6 OCPUs

### Materials

* This Guide
* Data
  + us\_states.csv

## ACCS – Tweet Producer

## Kafka – Event Hub

## BDC – Spark Streaming & SQL

## OAC – Analytics

# Integrating Other Data Sources

# References

Jernej Kase. “*Using TensorFlow on BDCS-CE with a Bootstrap script*”. <https://blogs.oracle.com/emeapartnerbiepm/using-tensorflow-on-bdcs-ce-with-a-bootstrap-script>

David Bayard. *“Big Data Journey: New Data Lake Workshop”.* <https://github.com/oracle/learning-library/tree/master/workshops/journey2-new-data-lake>

## Service Versions

|  |  |
| --- | --- |
| **BDC** | **Version: 18.1.4-1801182142** |
| ALLUXIO | 1.2.0 |
| AMBARI\_METRICS | 0.1.0 |
| HDFS | 2.7.1.2.4 |
| HIVE | 1.2.1.2.4 |
| LOGSTASH | 0.0.1 |
| MAPREDUCE2 | 2.7.1.2.4 |
| NGINX | 0.0.1 |
| PIG | 0.15.0.2.4 |
| SPARK2 | 2.1.0 |
| SPOC-UI | 0.5.0 |
| SPOCS\_FABRIC\_SERVICE | 0.1 |
| TEZ | 0.7.0.2.4 |
| YARN | 2.7.1.2.4 |
| ZEPPELIN | 0.7.0 |
| ZOOKEEPER | 3.4.6.2.4 |
| **EHCS** | **Version: 0.10.2.1.5-132** |
| Kafka | 0.10.2.1.5-132 |
| **OAC** | **Version: 18.1.3-1279** |