**Cluster selection**

1. Used Kafka on Azure HDInsight (Version: Kafka 1.0.0 (HDI 3.6))
2. Used Spark for stream processing on Azure HDInsight (Version: Spark 2.3.0 (HDI 3.6))

**Component for reading the data**

1. **FileName: EventGenerator.py**
2. Using kafka-python package to send the events to kafka
3. Created a virtual network between local system and the cluster. This is required as Azure HDInsight does not accept connection from public IPs
   1. I have used this architecture just to make sure that component sending the data and processing of events should happen on separate machines

**Event Processing**

1. Used Jupyter notebook to do event processing on spark
   1. **FileName: AlertGeneration.ipynb**

**Architecture**

1. **rawevents:** Data from household.csv has been moved to “rawevents” topic using EventGenerator.py script
2. **streamPerHourPerDay:** Created a query which will sum the consumption across house\_id, household\_id, Date, Lasthour and move it to “streamPerHourPerDay” topic
3. **streamPerHour**: Created another query which will take the mean and SD of the consumption across house\_id, household\_id, Lasthour and move it to “streamPerHour” topic
4. **streamPerHourAlert2:** Created another query which will take the mean and SD of the consumption across house\_id, Lasthour and move it to “streamPerHourAlert2” topic
5. **alertType1query:** Perform the **left join** betweenthe output of **streamPerHourPerDay** and **streamPerHour** topics and generate the CSVs at /example/AlertType1
6. **alertType2query:** Perform the **left join** betweenthe output of **streamPerHourPerDay** and **streamPerHourAlert2** topics and generate the CSVs at /example/AlertType2
7. Merge the files generated at /example/AlertType1 into 1 file
8. Merge the files generated at /example/AlertType2 into 1 file