Seminar Report on "Big Data PySpark"

1. Date of the Seminar/Workshop: 19-11-2020

2. Title of the Seminar/Workshop: "Big Data PySpark"

3. Name of the Speaker/Resource person with Affiliation with the institute/industry:

Mr. Aaron Rebello

Industry:- Capgemini Technology Services India Limited

Designation:-Associate Consultant

Role:-Big Data Engineer

4. Venue of the Seminar/Workshop: Platform used-Google meet

5. Duration of the Seminar: 3 hrs(9.30 am to 12.30 pm)

6. Conducted For: Students of final year computer Engineering

7. Objective of the Seminar/Workshop /Curriculum Gap identified/Other than that:

The objective of the Seminar was basically to ensure that the students should have a clear understanding of big data and how to start projects in this domain as students have to implement mini project based on this subject. And as a speaker is working in industry as a Big Data Engineer so the students should take benefit and a lot of insights could be inferred from his experience in this industry

8. Contents of the Seminar/Workshop:

The 3 hours period was very interactive and aroused key enthusiasm among students. The speaker started with

- Data timeline starting from the 1960s to the present time.
- OLTP vs. OLAP
- OLAP architecture
- ETL tools such as Xplenty, Stitch, Informatica, AWS and Oracle
- Different tools of Big Data like Haoop,Kafka,Spark,Storm
- The working of MapReduce
- The spark ecosystem
- RDD- the resilient distributed dataset

and finally hands-on/demonstration of basic query structures using cloudera and pyspark

9. Description of the Entire Event

The Computer Department of St. Francis Institute of Technology organized a seminar on "Big Data PySpark" on Thursday 19th November-2020 between 9.30 am to 12 noon , online on Google meet.

There were approximately 80 student participants and 4 faculty participants who attended the seminar. The program was started with welcome speech by Ms. Anuradha Srinivasaraghavan where she introduced the Speaker. Later, Mr Aaron Rebello took over the session.

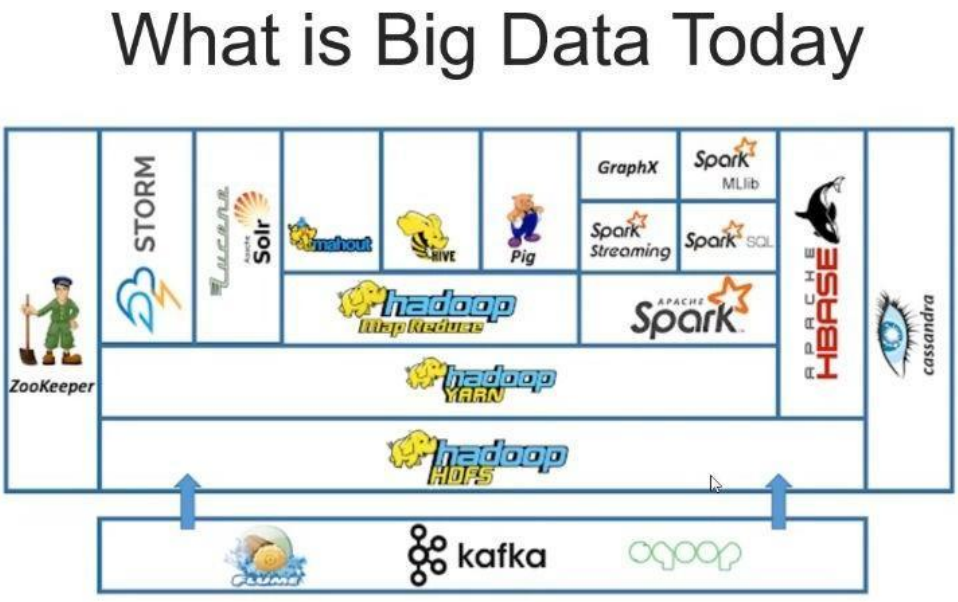
The speaker started off by introducing himself and his experience in the domain followed by the projects taken up in the organization he works in.

Then he started off with the main topic of the seminar by first talking about the data timeline starting from the 1960s to the present time. He spoke about how in 1960 database was introduced and then in 1970 OLTP came into the picture, followed by OLAP in 1975 and big data in 2005. Then a brief introduction was given about the OLAP architecture and juxtaposed OLAP with OLTP so as to understand the importance of OLAP. Then ETL tools were introduced to us like Xplenty, Stitch, Informatica, AWS and Oracle. He talked about the data distribution pyramid and how there is a very less amount of structured data and a large amount of unstructured data.

Furthermore, he told us about the emerging of big data due to the three papers from Google:

1) Google file system (2003)
2) MapReduce (2004)
3) BigTable (2006)

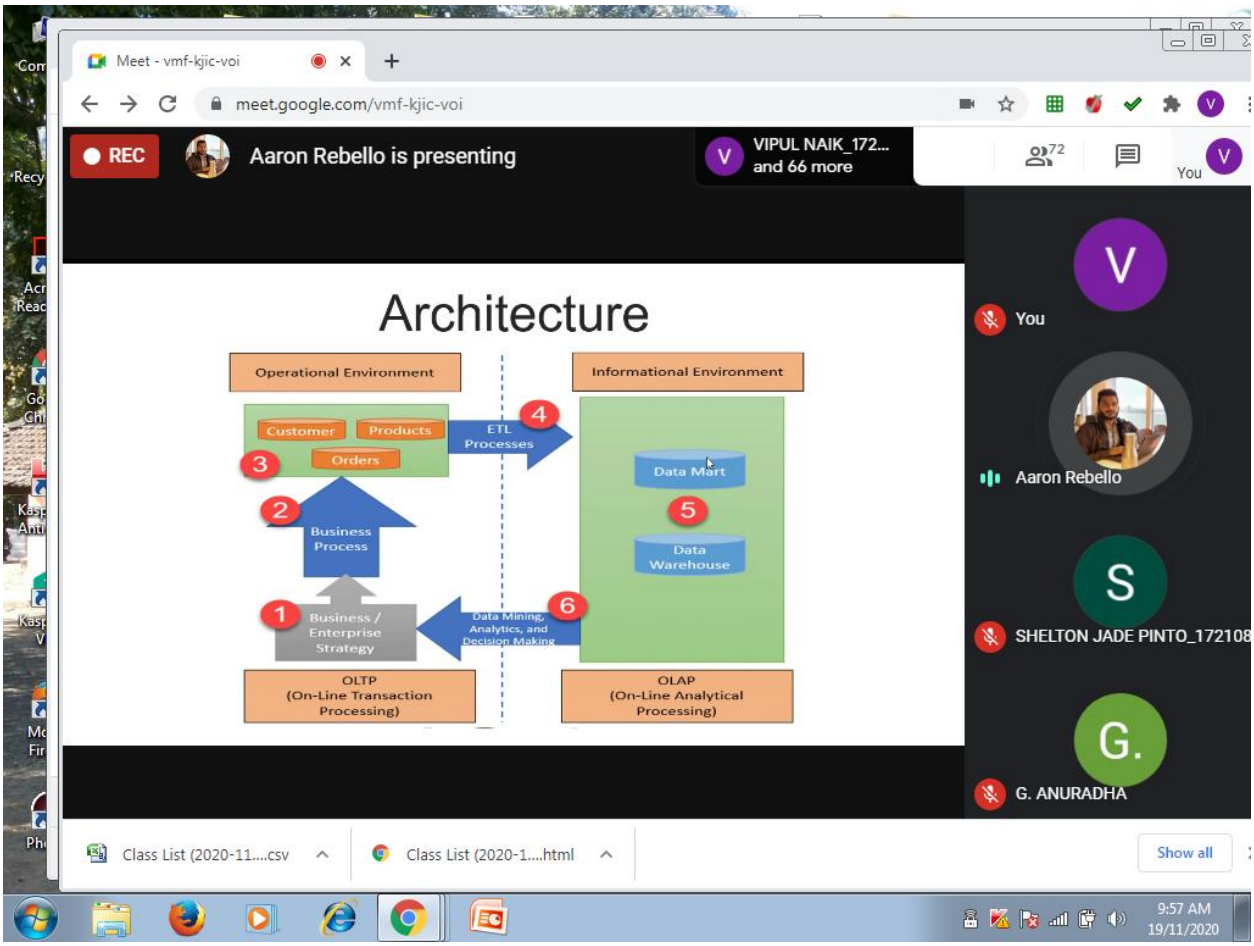
Then the definition of big data was introduced to us along with its different tools as shown below:

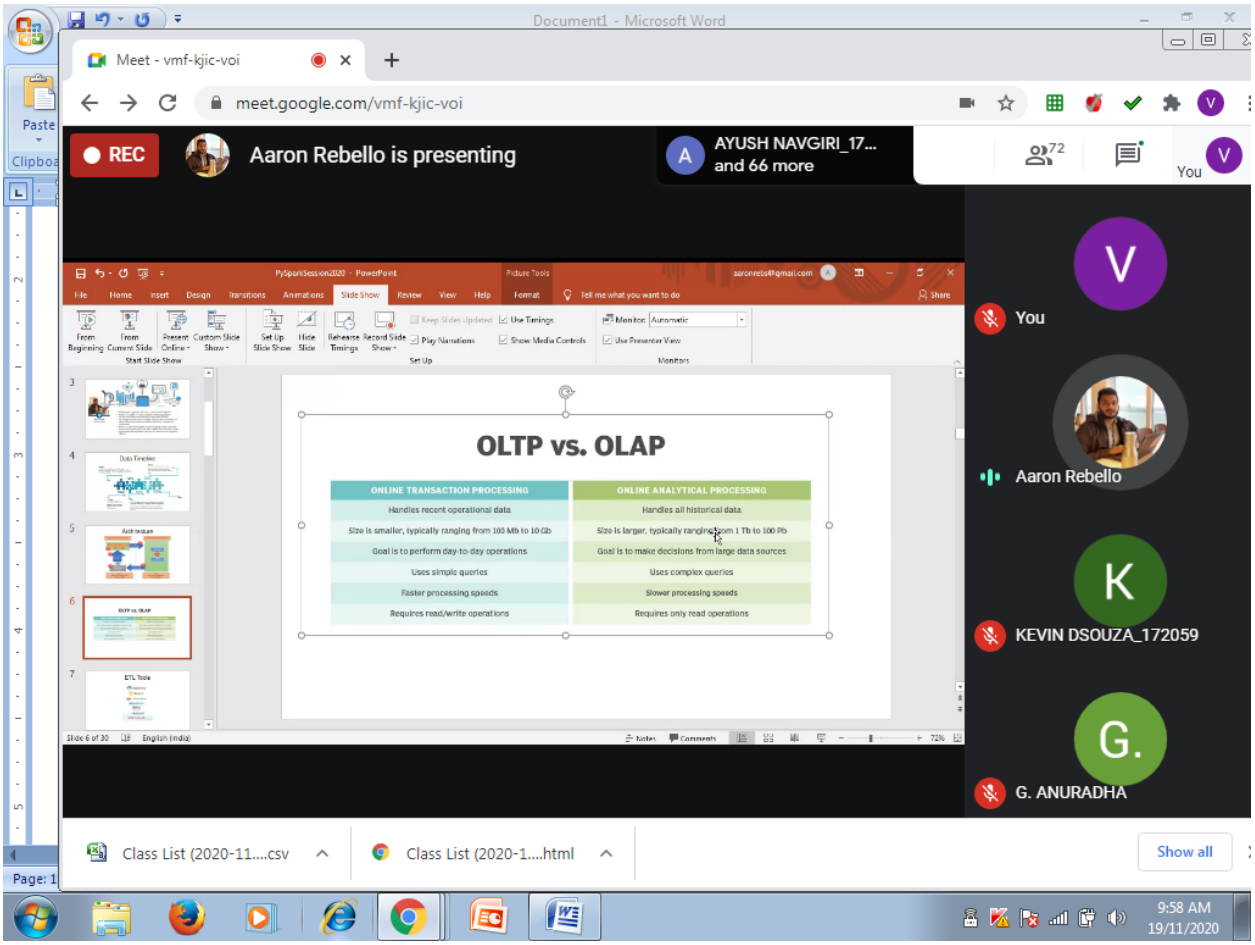


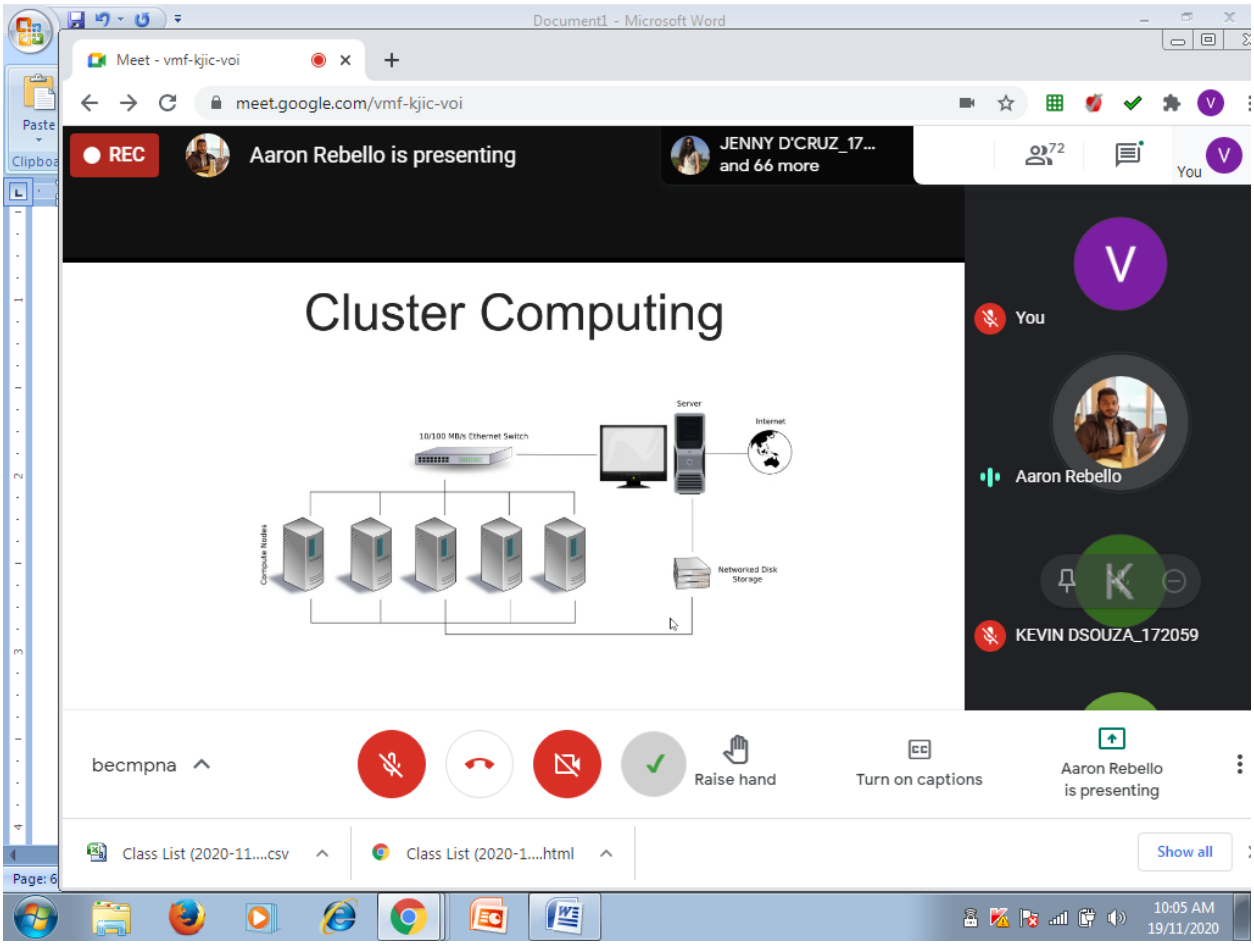
Moving further, he explained the working of MapReduce in detail. MapReduce works in 2 phases. In the first phase, mapping is done where the jobs are split and data is mapped. In the second phase, the reduce task shuffles and consolidates the data. But this process is really slow because data replication takes place. Therefore, Spark came into the picture since it is 10-100 times faster. The spark ecosystem was explained further. It has service APIs like Spark Streaming, Spark SQL and GraphX, computing engine - Spark core, resource manager - yarn and Mesos and distributed file system = HDFS and HBase. Then he told us about RDD which is the resilient distributed dataset and how spark relies on this as it can be stored in memory and also perform various operations. Then he went on to show the code and work in the Java and Python languages.

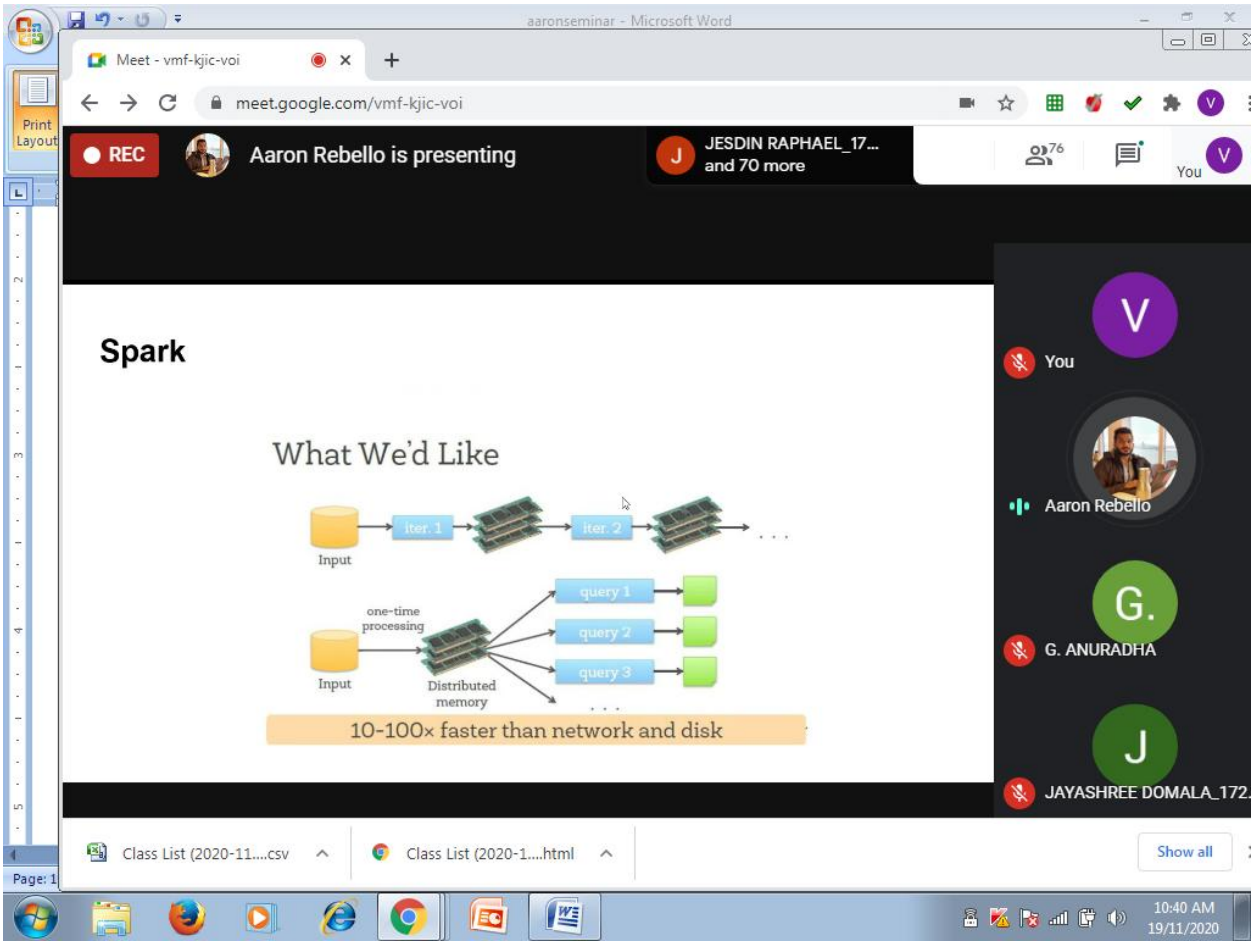
At the end of the session, the forum was open for doubts where he solved the queries posed by the students. He got us acquainted with the set up of the system so as to be able to work on big data projects. He also told us about the resources available on the internet like the official documentation which can be referred for practical implementation. Due to the session, the students had a clearer understanding of big data and how to start projects in this domain. It was very beneficial and a lot of insights could be inferred from his experience in this industry.

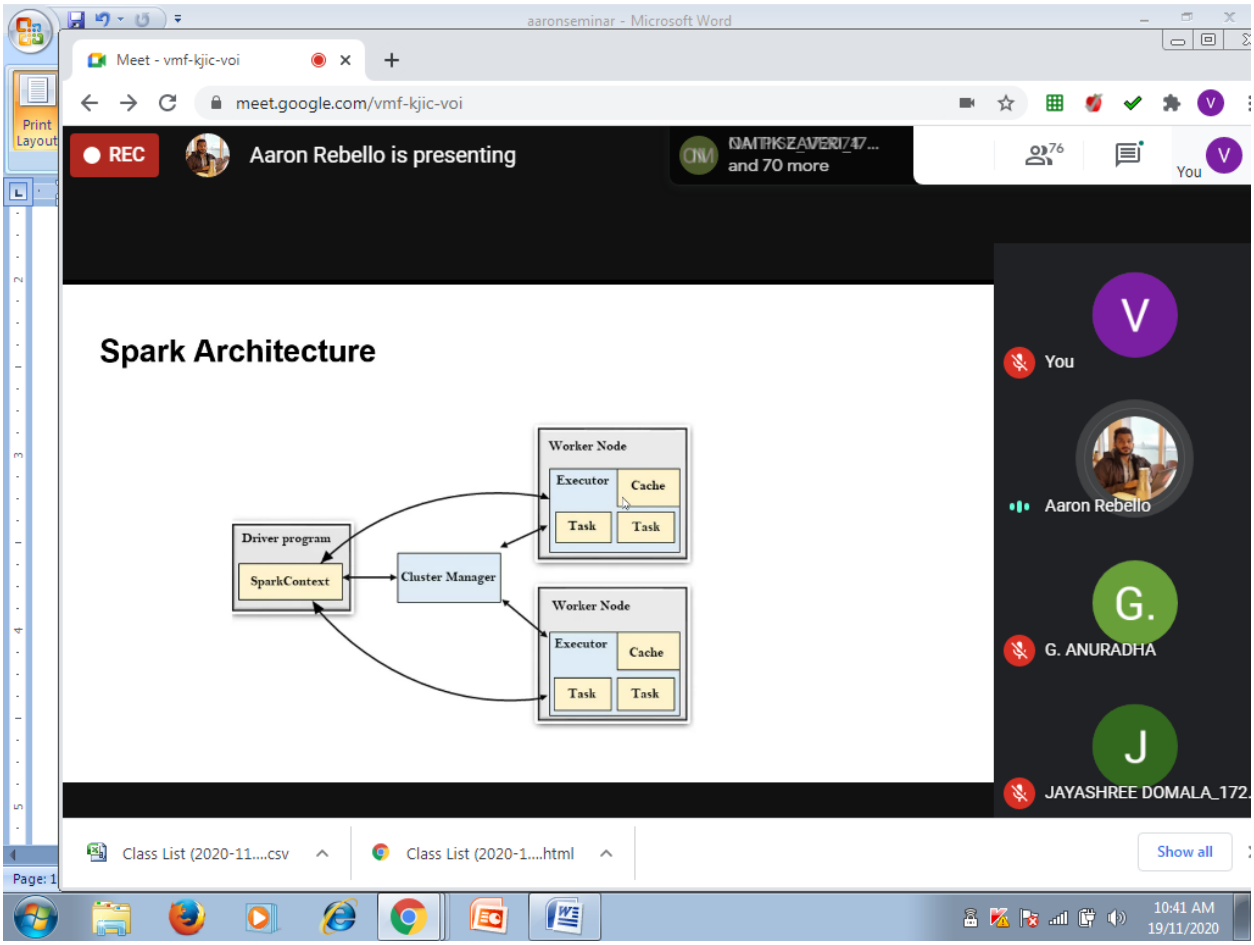
Session
Photographs

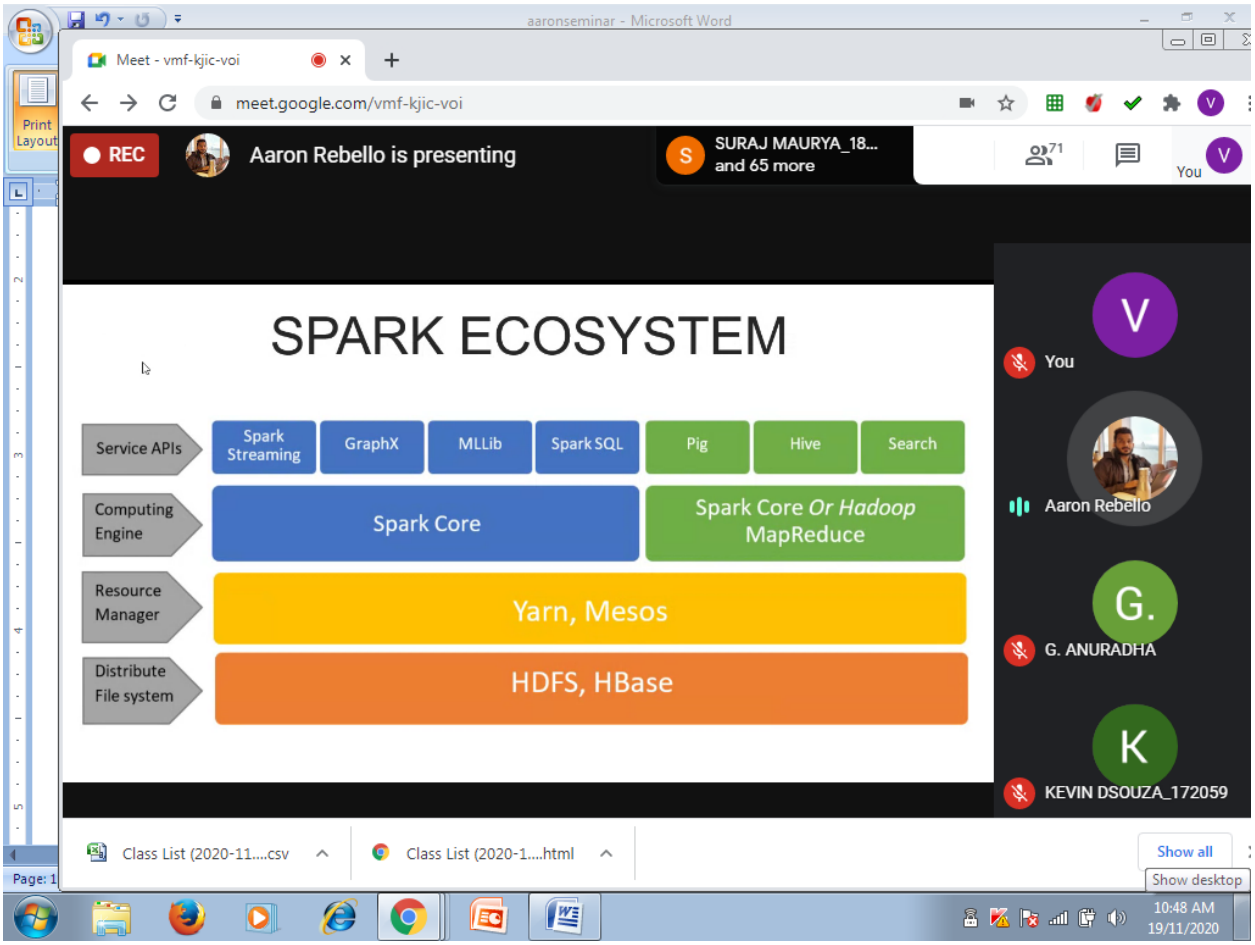


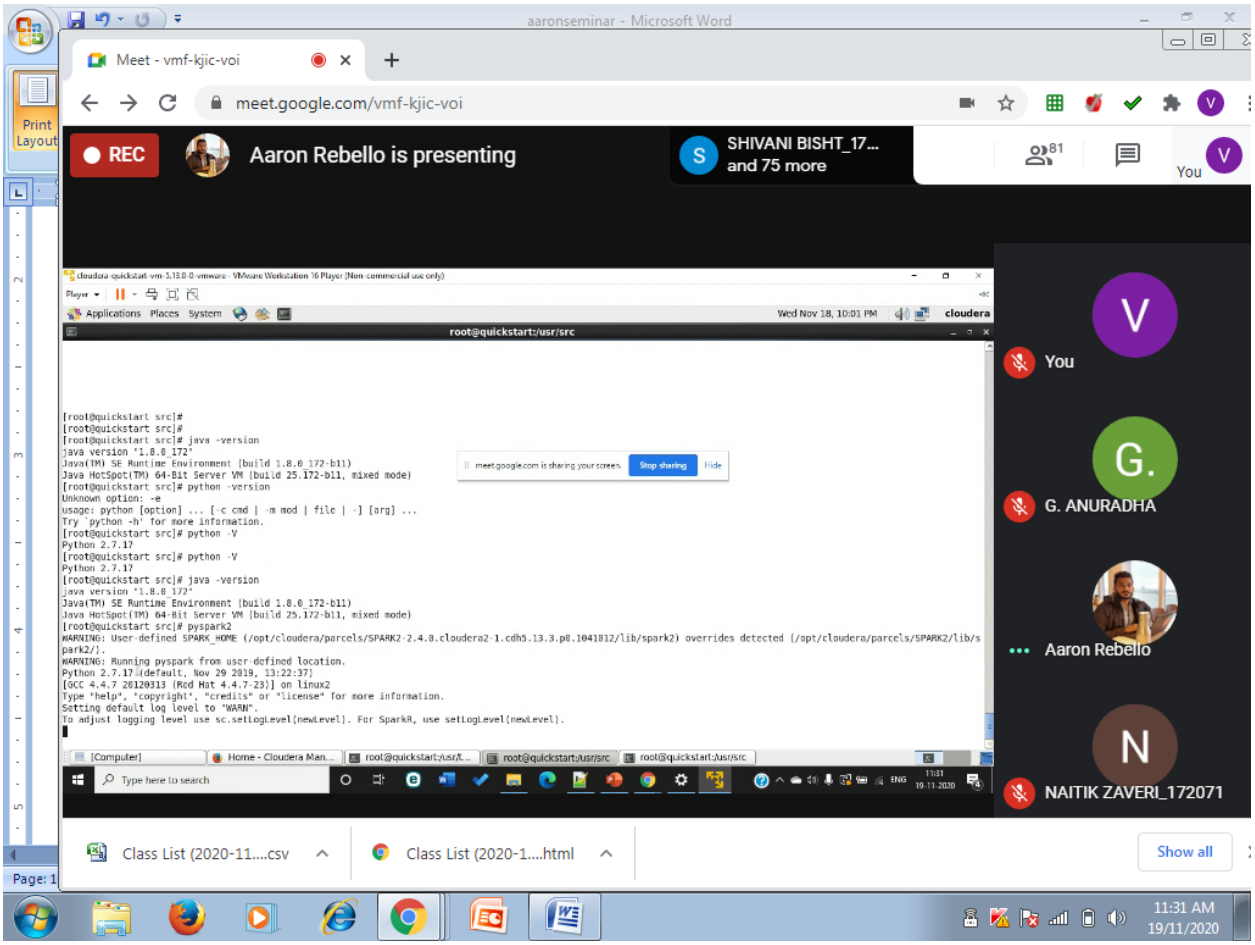

Document1 - Microsoft Word
Meet - vmf-kjic-voi
meet.google.com/vmf-kjic-voi
REC
Aaron Rebello is presenting
AYUSH NAVGIRI_17...
and 66 more
You
OLTP vs. OLAP
ONLINE TRANSACTION PROCESSING
ONLINE ANALYTICAL PROCESSING
Handles recent operational data
Handles all historical data
Size is smaller, typically ranging from 100 Mb to 10 Gb
Size is larger, typically ranging from 1 Tb to 100 Pb
Goal is to perform day-to-day operations
Goal is to make decisions from large data sources
Uses simple queries
Uses complex queries
Faster processing speeds
Slower processing speeds
Requires read/write operations
Requires only read operations
You
Aaron Rebello
KEVIN DSOUZA_172059
G. ANURADHA
Class List (2020-11....csv
Class List (2020-1....html
Show all
9:58 AM
19/11/2020

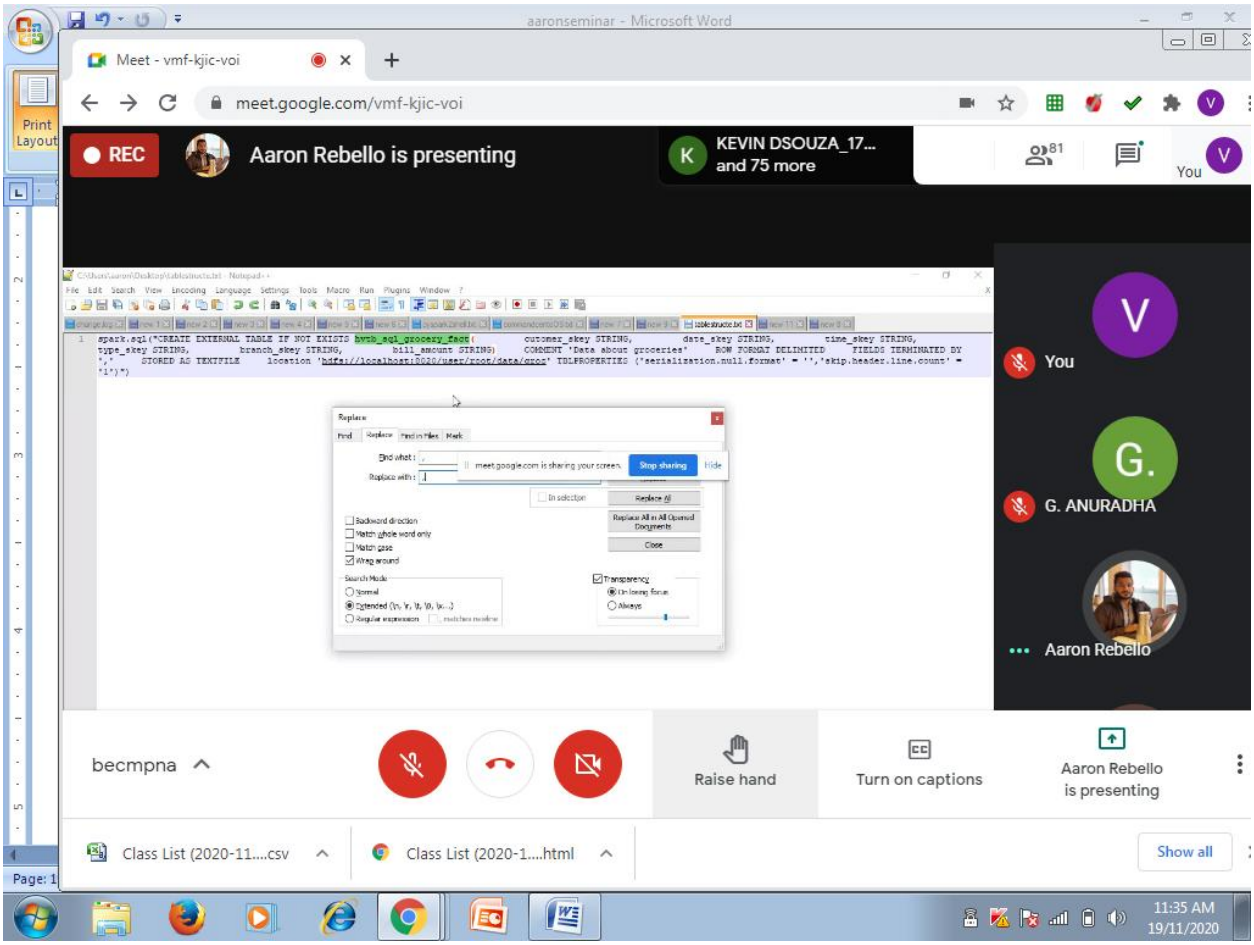

Document1 - Microsoft Word
Meet - vmf-kjic-voi
meet.google.com/vmf-kjic-voi
REC
Aaron Rebello is presenting
JENNY D'CRUZ_17...
and 66 more
You
Cluster Computing
10/100 MB/s Ethernet Switch
Server
Internet
Compute Nodes
Networked Disk Storage
You
Aaron Rebello
KEVIN DSOUZA_172059
becmpna
Raise hand
Turn on captions
Aaron Rebello is presenting
Class List (2020-11....csv
Class List (2020-1....html
Show all
10:05 AM
19/11/2020


aaronseminar - Microsoft Word
Meet - vmf-kjic-voi
meet.google.com/vmf-kjic-voi
REC
Aaron Rebello is presenting
JESDIN RAPHAEL_17...
and 70 more
You
Spark
What We'd Like
Input
iter. 1
iter. 2
one-time
processing
Input
Distributed
memory
query 1
query 2
query 3
10-100× faster than network and disk
You
Aaron Rebello
G. ANURADHA
JAYASHREE DOMALA_172...
Class List (2020-11....csv
Class List (2020-1....html
Show all
10:40 AM
19/11/2020


aaronseminar - Microsoft Word
Meet - vmf-kjic-voi
meet.google.com/vmf-kjic-voi
REC
Aaron Rebello is presenting
and 70 more
You
Spark Architecture
Worker Node
Executor
Cache
Task
Task
Driver program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task
You
Aaron Rebello
G.
G. ANURADHA
J
Class List (2020-11....csv
Class List (2020-1....html
Show all
10:41 AM
19/11/2020


aaronseminar - Microsoft Word
Meet - vmf-kjic-voi
meet.google.com/vmf-kjic-voi
REC
Aaron Rebello is presenting
SURAJ MAURYA_18...
and 65 more
71
You
SPARK ECOSYSTEM
Service APIs
Spark Streaming
GraphX
MLLib
Spark SQL
Pig
Hive
Search
Computing Engine
Spark Core
Spark Core Or Hadoop MapReduce
Resource Manager
Yarn, Mesos
Distribute File system
HDFS, HBase
You
Aaron Rebello
G. ANURADHA
KEVIN DSOUZA_172059
Class List (2020-11....csv
Class List (2020-1....html
Show all
Show desktop
10:48 AM
19/11/2020


aaronseminar - Microsoft Word
Meet - vmf-kjic-voi
meet.google.com/vmf-kjic-voi
REC
Aaron Rebello is presenting
SHIVANI BISHT_17...
and 75 more
You
root@quickstart:/usr/src
[root@quickstart src]#
[root@quickstart src]#
[root@quickstart src]# java -version
java version "1.8.0_172"
Java(TM) SE Runtime Environment (build 1.8.0_172-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.172-b11, mixed mode)
[root@quickstart src]# python -version
Unknown option: -e
usage: python [option] ... [-c cmd | -m mod | file | -] [arg] ...
Try `python -h' for more information.
[root@quickstart src]# python -V
Python 2.7.17
[root@quickstart src]# python -V
Python 2.7.17
[root@quickstart src]# java -version
java version "1.8.0_172"
Java(TM) SE Runtime Environment (build 1.8.0_172-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.172-b11, mixed mode)
[root@quickstart src]# pyspark2
WARNING: User-defined SPARK_HOME (/opt/cloudera/parcels/SPARK2-2.4.0.cloudera2-1.cdh5.13.3.p0.1041812/lib/spark2) overrides detected (/opt/cloudera/parcels/SPARK2/lib/spark2/).
WARNING: Running pyspark from user-defined location.
Python 2.7.17 (default, Nov 29 2019, 13:22:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-23)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
meet.google.com is sharing your screen.
Stop sharing
Hide
You
G. ANURADHA
Aaron Rebello
NAITIK ZAVERI_172071
Class List (2020-11....csv
Class List (2020-1....html
Show all
11:31 AM
19/11/2020



Ms. Varsha N and Ms. Snehal K. **Dr. Kavita Sonawane**

Seminar Incharge **HOD, CMPN**