Case Study

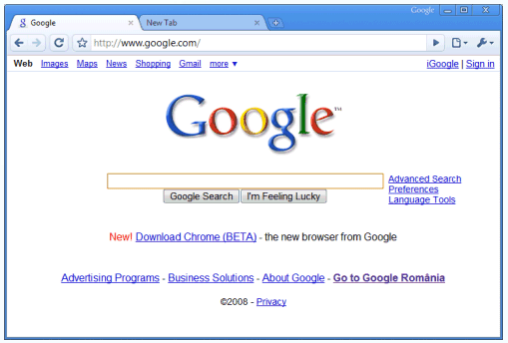
-

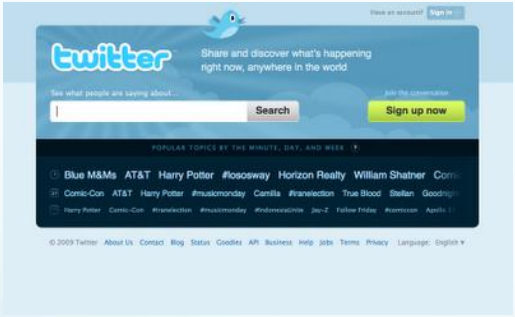
WebLog Analysis through Hadoop

Objective

- How Hadoop and it’s stack component interact.
- How to solve the business use case using huge amount of data.
- How large scale processing performed.
- How to efficiently manage BigData.

BigData





- Huge amount of data that is in TB's / PB's
- Traditional Data management technologies can't handle this data.
- 3 Dimensions - Velocity, Volume and Variety

Hadoop

- Hadoop is distributed computing framework to solve the BigData problems.
  - **Scalable:** store and process petabytes, scale by adding Hardware
  - **Economical:** 1000's of commodity machines
  - **Efficient:** run tasks where data is located
  - **Reliable:** data is replicated, failed tasks are rerun

- Provides many component like Hive, Pig, HBase, Sqoop, Flume, Oozie etc.

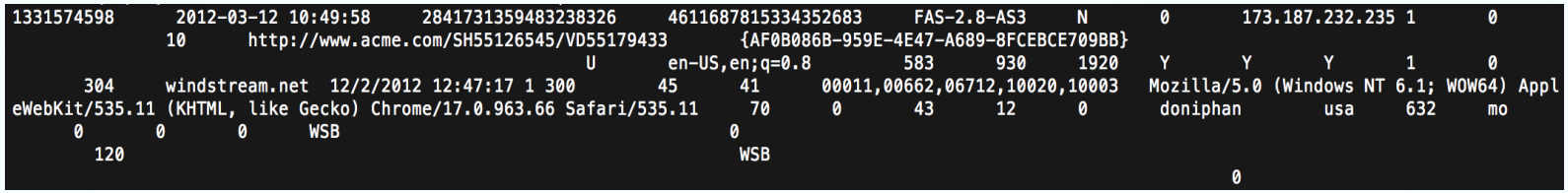
WebLog Analysis

- Analyze weblogs to solve following business use case:
  - What products do visitor like location wise and what are they most likely to buy in future.

- To solve this, we require weblogs along with product and customer data.
  - **Clickstream logs:** Website log files containing information such as URL, timestamps, IP address, geo-coded IP address, and user ID (SWID).
  - **User Log:** User data listing SWIDs (Software User IDs) along with date of birth and gender.
  - **Product logs:** Product data that maps product categories to website URLs.

Data

- Clickstream logs will be in form of files and periodically available on FS



- Users and Products will be in form of table and will be initially available on MySQL.

SWID	BIRTH_DT	GENDER_CD
0001BDD9-EABF-4D0D-81BD-D9EABFCD0D7D	8-Apr-84	F

url	category
http://www.acme.com/	books

Application Flow

- Copy periodically generated Clickstream logs on HDFS.
- Copy products table on Hive using Sqoop import command from MySQL.
- Copy users table on Hive using Sqoop import command from MySQL.
- Execute MapReduce on clickstream logs to generate the required fields (like timestamp, ip, url, swid, city, country, state) and populate the output into partitioned Hive table.
- Execute Hive script to generate the final data using partitioned filtered web data, products and users Hive table.
- Copy the final processed output data to MySQL using Sqoop export command for data analysis.

Step 1

- Copy web logs on HDFS using DFS put or Flume in some datestamp directory classified on datestamp basis.

  $ hadoop dfs –put /bigdata/weblogs/20160604 hdfs://HadoopMaster:8020/user/root/weblogs/20160604

Step 2

- Create the external partitioned Hive table weblogs to store the filtered fields from stored logs file on HDFS.

  - hive > CREATE EXTERNAL TABLE weblogs

    (ts STRING, ip STRING, url STRING, swid STRING, city STRING, country STRING, state STRING)

    PARTITIONED BY (dt STRING)

    ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t’

    STORED AS TEXTFILE;

Step 3

- Copy Products table on Hive through Sqoop import command.

  - hive > CREATE TABLE products

    (url STRING, category STRING)

    ROW FORMAT DELIMITED FIELDS TERMINATED by '\001'
    STORED AS TEXTFILE

    TBLPROPERTIES ("skip.header.line.count"="1");

  - sqoop import --connect jdbc:mysql://localhost:3306/mydb --driver com.mysql.jdbc.Driver --username root --password root123 --table products --hive-import -m 1

Step 4

- Copy users table on Hive using Sqoop import command

  - hive > CREATE TABLE users
    (swid STRING, birth_dt STRING, gender_cd CHAR(1))
    ROW FORMAT DELIMITED FIELDS TERMINATED by '\001’
    STORED AS TEXTFILE
    TBLPROPERTIES ("skip.header.line.count"="1");

  - sqoop import --connect jdbc:mysql://localhost:3306/mydb --driver com.mysql.jdbc.Driver --username root --password root123 --table users --hive-import -m 1

Step 5

- Develop and execute the MapReduce program on weblogs to generate the required fields.
  - $ hadoop jar webloganalysis-1.0-SNAPSHOT.jar

    org.examples.FilterWebLogs

    -Dfields=1,7,12,13,49,50,52

    /user/root/weblogs/20160604

    /user/root/demographic/20160604

  - Add the generated filtered data as partitioned Hive table.

    - hive > ALTER TABLE weblogs

      ADD PARTITION (dt='20160604')

      LOCATION ‘/user/root/demographic/20160604'

    - ';

Step 6

- Now execute the webloganalytics hive script to generate the final data.

  - hive > CREATE TABLE webloganalytics AS SELECT

    to_date(w.ts) logdate, w.url, w.ip, w.city, upper(w.state) state, w.country, p.category, CAST(datediff(from_unixtime(unix_timestamp()), from_unixtime(unix_timestamp(u.birth_dt, 'dd-MMM-yy'))) / 365 AS INT) age, u.gender_cd FROM weblogs w

    INNER JOIN products p on w.url = p.url AND AND w.dt='20160604'

    LEFT OUTER JOIN users u on w.swid = concat('{', u.swid , '}');

Step 6

- Now execute the webloganalytics hive script to generate the final data.

  - hive > CREATE TABLE webloganalytics AS SELECT

    to_date(w.ts) logdate, w.url, w.ip, w.city, upper(w.state) state, w.country, p.category, CAST(datediff(from_unixtime(unix_timestamp()), from_unixtime(unix_timestamp(u.birth_dt, 'dd-MMM-yy'))) / 365 AS INT) age, u.gender_cd FROM weblogs w

    INNER JOIN products p on w.url = p.url AND AND w.dt='20160604'

    LEFT OUTER JOIN users u on w.swid = concat('{', u.swid , '}');

Step 7

- Copy the final processed output data to MySQL using Sqoop export command.

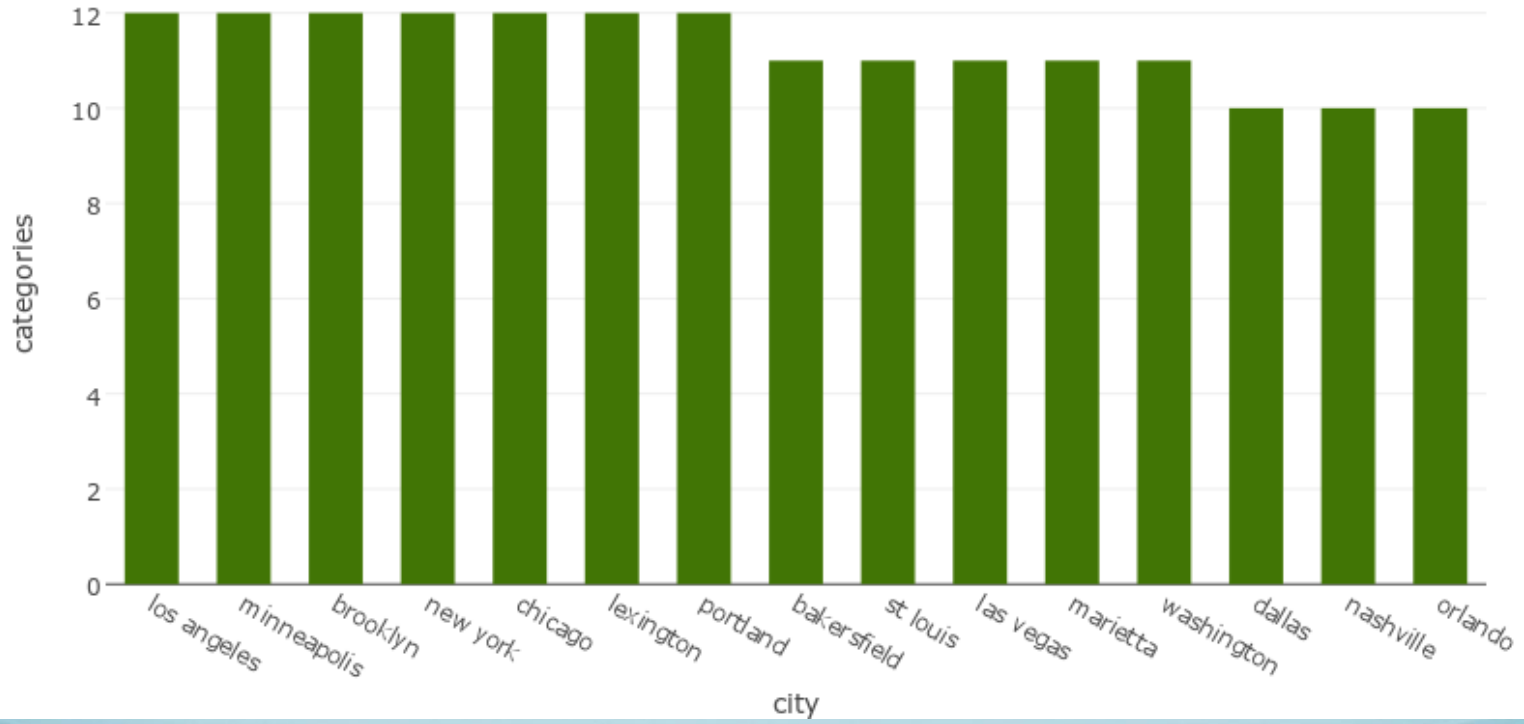
  - mysql> CREATE TABLE webloganalytics (logdate VARCHAR(100),
    URL VARCHAR(1000), ip VARCHAR(40), city VARCHAR(50),
    state VARCHAR(30), country VARCHAR(30), category VARCHAR(30),
    age VARCHAR(5), gender_cd CHAR(5));

  - $ sqoop export --connect jdbc:mysql://localhost:3306/mydb --username root --password root123 --table webloganalytics--export-dir /user/hive/warehouse/webloganalytics --input-fields-terminated-by '\001'

Data Analysis

- To perform the data analysis on processed data following queries can be executed on MySQL.
  - SELECT city, COUNT(DISTINCT category) as category FROM webloganalytics GROUP BY (city) ORDER BY category desc limit 15;
  - SELECT city, GROUP_CONCAT(DISTINCT category) AS category FROM webloganalytics GROUP BY(city);
  - SELECT age, GROUP_CONCAT(DISTINCT category) AS category FROM webloganalytics GROUP BY (age);
  - SELECT city, GROUP_CONCAT(DISTINCT category) AS category, GROUP_CONCAT(DISTINCT gender_cd) AS gender_cd FROM webloganalytics GROUP BY(city);

UI View of processed data



Next

- Use Oozie to create Workflow and schedule it through Coordinator.
- Use of Flume how to transfer logs file.
- Overview of Hortonworks Data Platform.