**Technical Report with Detailed Code Appendix**

**Patent Data Extraction**

**- Phase 1**

**Team Kick SAS**

**Members:**

**Atluri Laxmi Narayana 800907065**

**Mohan Kumar Nallapareddy 800886720**

**Punniya Dharshan Ganesan 800894568**

**Reethu Gundavarapu 800903616**

**Sharan Kumar Reddy Kadire 800888186**

**Sri Harsha Degala 800902540**

**TECHNICAL SUMMARY**

**DATA EXTRACTION:**

***Tools used: Hadoop-Hive, Schell Script.***

Data Extraction is achieved by following the follow steps:

1. To use a Schell script to extract Microsoft patents from cluster.
2. Run a Hive code on the extracted data to extract the desired columns.
3. Copy the contents to a single spreadsheet.

**Step 1:** To extract Microsoft patents from cluster:

Below is the Schell script Code we used on “Bibliodata.txt”

#!/in/sh

 find -L dsba-6100/patentData2000\_2015/patBiblio2000\_1h2015 |grep -i ".zip" >Bibliodata.txt

 path=""

lines=$(cat Bibliodata.txt|wc -l)

echo $lines

for (( a=1; a<=$lines;a++ ))

do

echo "extracting file $a of $lines "

path=$(sed -n "$a"p Bibliodata.txt )

unzip -u $path -d unzip\_biblio/

done

awk -v RS="</us-patent" '/<orgname>Microsoft/{print}' unzip\_biblio/ ipgb\*.xml >> ipgb.xml.txt

* This script extracts the Raw files by unzipping the ipgb files from patBiblio folder in the cluster.
* AWK command used in the script “awk -v RS="</us-patent" '/<orgname>Microsoft/{print}' unzip\_biblio/ ipgb\*.xml > ipgb.xml.txt” extracts patent data related to Microsoft.

**Step 2:**

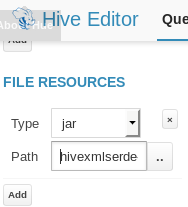
Ran Hive code on the extracted data to extract the desired columns .Once we have the Output file (obtained by running the script in particular because of the AWK command)

*Software setup:*

For easy access to Hadoop we installed Cloudera 5.4.2.0.

We installed Oracle Virtual box to work with Cloudera.

After Successful installation of the above software’s. We need to place the jar file **hivexmlserde-1.0.5.3.jar** in Hive environment. This class is used to work with XML files.



***Placing the Jar file in Hive Editor****.*

***CODE:***

***Creating Table to load XML data:***

*create table q\_patent2*

*(file string)*

*ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'*

*STORED AS*

*INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlSerDe'*

*OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'*

*TBLPROPERTIES (*

*"xmlinput.start"="<us-patent-grant",*

*"xmloutput.end"="</ us-patent-grant >"*

*);*

-------------------------------------------------------------------------------------------------------------------------------------*load data inpath 'hdfs:/user/hive/ipgb.xml' into TABLE patent\_ipgb* --------------------------------------------------------------------------------------------------------------------------------------

***Text parsing to get patent\_no, date-publ, date-produced:***

*select substr(*

*file,*

*(instr(file,'le="'))+4,*

*(instr(file,'.XML'))-(instr(file,'file='))-6)*

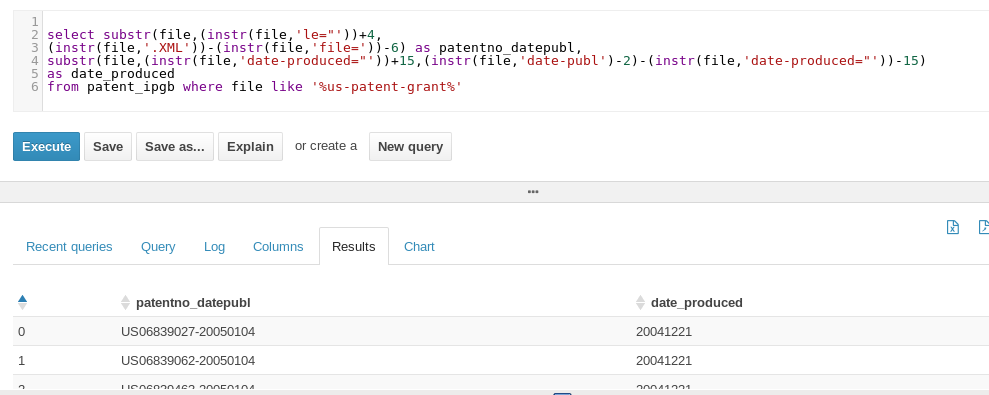
*as patentno\_datepubl,*

*substr(file,(instr(file,'date-produced="'))+15,*

*(instr(file,'date-publ')-2)-(instr(file,'date-produced="'))-15)*

*as date\_produced*

*from patent\_ipgb where file like '%us-patent-grant%'*



***patentno\_datepubl, date\_produced fields are extracted using Hive.***

***Text parsing to get invention-title:***

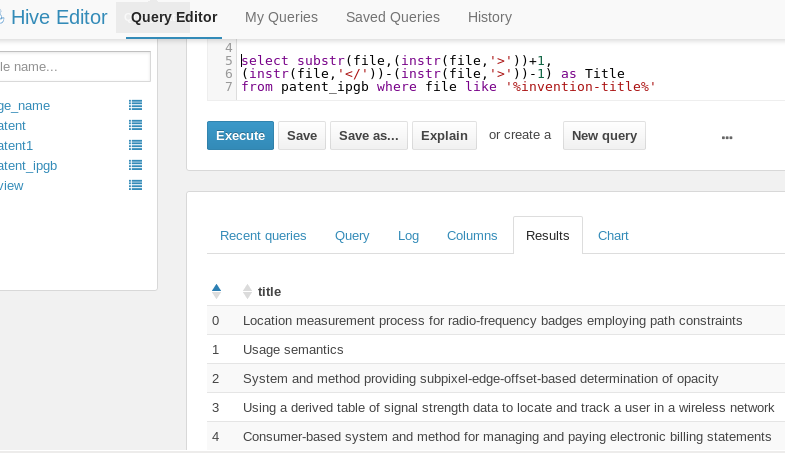
*INSERT OVERWRITE DIRECTORY 'hdfs:/user/latluri'*

*select substr(file,(instr(file,'>'))+1,*

*(instr(file,'</'))-(instr(file,'>'))-1)*

*as Title*

*from apple\_title where file like '%invention-title%'*

**

***Invention-title field is extracted using Hive.***

***Text Parsing to get abstract:***

*INSERT OVERWRITE DIRECTORY 'hdfs:/user/latluri'*

*select substr*

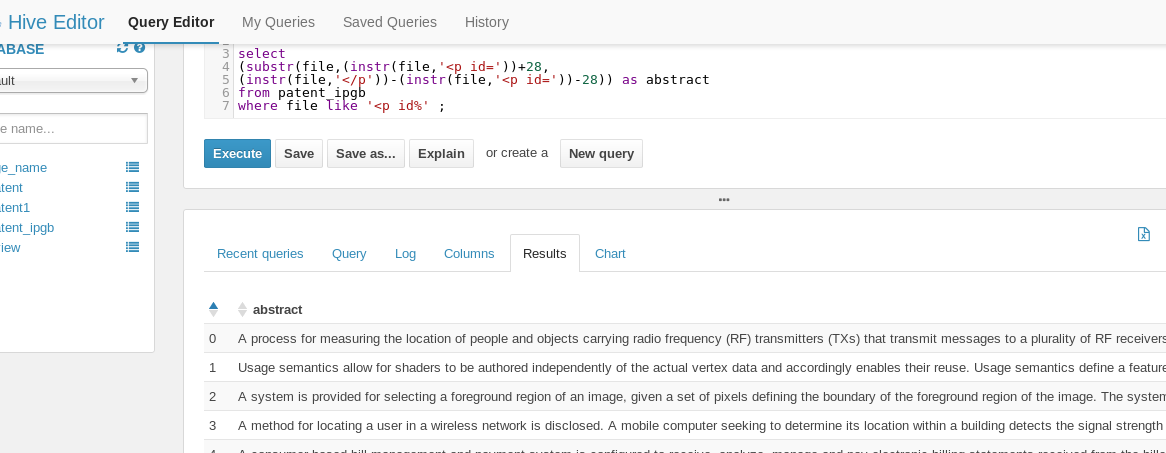
*(file,(instr(file,'<p id='))+28,*

*(instr(file,'</p'))-(instr(file,'<p id='))-28)*

*as abstract*

*from apple\_title*

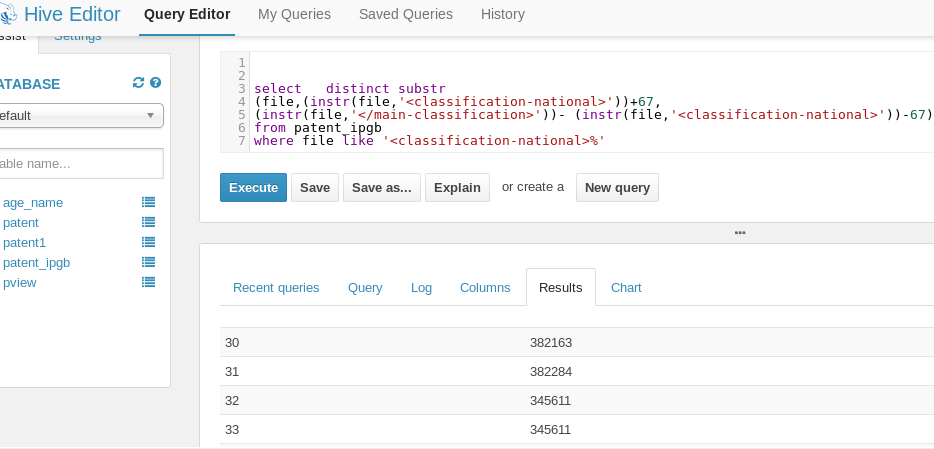
*where file like '<p id%' ;*

**

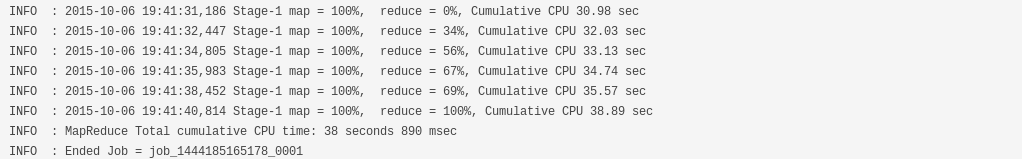
***Abstract field is extracted using Hive.***

***Text Parsing to get Class:***

*select  distinct substr  
(file,(instr(file,'<classification-national>'))+67,  
(instr(file,'</main-classification>'))- (instr(file,'<classification-national>'))-67)  
from patent\_ipgb   
where file like '<classification-national>%';*

**

***Log showing Map/Reducer Function in Hive:***



**DATA VISUALIZATION:**

***Tools used:*** R and Tableau

We placed the file in the source drive

Rplot02 - initial data

Rplot01\_th - Stop words

rplot - lower - lower

Imported the file into R:

> patent.txt <- VCorpus(DirSource("~/"))

<<PlainTextDocument>>

Metadata:  7

Content:  chars: 19671135

Ran wordcloud on the data:

> wordcloud(words = patent.txt,min.freq = 10,max.words = 500,random.order = 'false')

Ran wordcloud after removing the prepositions using stopwords:

> patent2.txt <- tm\_map(patent.txt,remove Words, stopwords('English'))

> wordcloud(words = patent2.txt,min.freq = 10,max.words = 500,random.order = 'false')

Ran wordcloud on title using colour:

>wordcloud(words=mypatenttxt3,colors=brewer.pal(8,’Dark2’),min\_freq = 10,max.words = 500,random.order = ‘false’)

* We imported the file in tableau and ran the visualizations.

The data Extracted:

