**Innovation Strategy for Microsoft**

**Summary**

**Data Wrangling Phase:**

We used Unix Shell Scripting to Unzip files and Hadoop-Hive for Text Mining. We have Shared the code for the same in *Appendix-1*.

**Analytics and Innovation Strategy:**

We tried to come up with a strategy by analyzing the patent data extracted in the Data Wrangling Phase.

We derived 4 factual columns from the Dimensional data we extracted from the Hadoop Cluster (code is attached in the *Code Appendix Phase-2*)

* No of forward citations
* No of Backward Citation
* Days for patent Granted
* No of Authors

**Strategy-1:**

**Bubble Chart Viz:**

Class-Database has high aging of Patents, though it is relatively having less patent count . This can contribute to revenue loss. *Refer Appendix-2. (Fig-1)*

**Regression:**

No of Authors per patent has a slight negative impact on Forward Citations. This in other words mean that increasing resource count in Database Class is not going to help . *Refer Appendix-2 (Fig-2 & 3)*

**Strategy-2:**

**Word clouds of Patent Abstract in R:**

Using the results of Word clouds we concurred that Microsoft's top patents-which we got it using our metrics is also areas where Microsoft is a Predominant Market Leader except for Mobile Market. *Refer Appendix-3 (Fig 4 & 6)*

**Topic Modeling:**

We aggregated the weights of each topic across all the documents obtained from result and used the weights of the words/topics to support the fact that it is same as the Word Cloud results. *Refer Appendix-3 (Table-1)*

**Appendix-1**

**Step 1:**

Below is the Schell script Code we used on to unzip Microsoft Data

*#!/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*

**Step 2:**

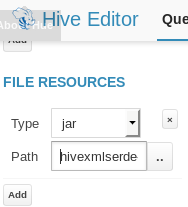
Ran Hive code on the extracted data to extract the desired columns .Once we have the Output file(obtained by running the script in particular because of the AWK command)

***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 patent\_ipgb*

*(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:***

*selectsubstr(*

*file,*

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

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

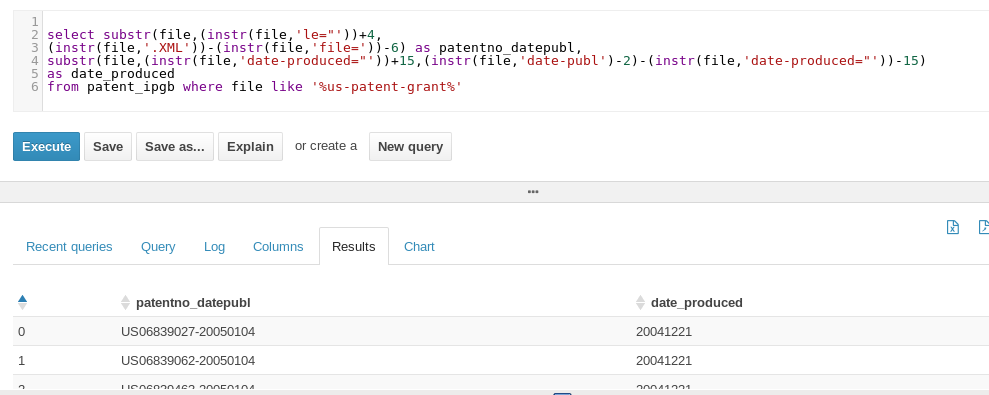
*aspatentno\_datepubl,*

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

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

*asdate\_produced*

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



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

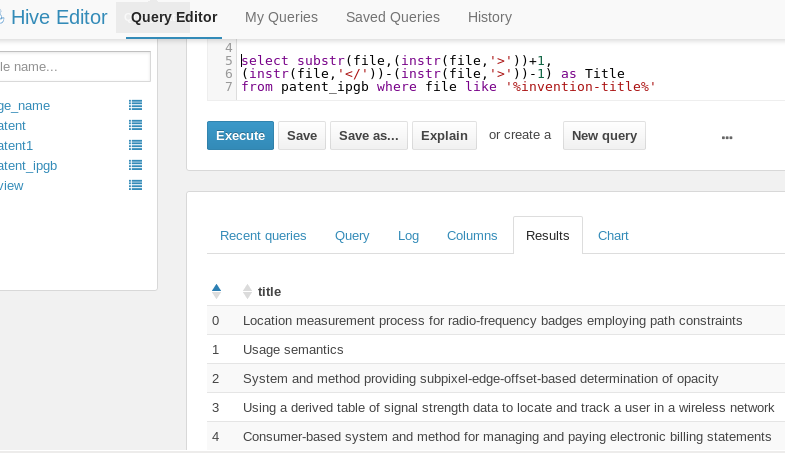
***Text parsing to get invention-title:***

*selectsubstr(file,(instr(file,'>'))+1,*

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

*as Title*

*frompatent\_ipgb where file like '%invention-title%'*

**

***Invention-title fieldis extracted using Hive.***

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

*selectsubstr*

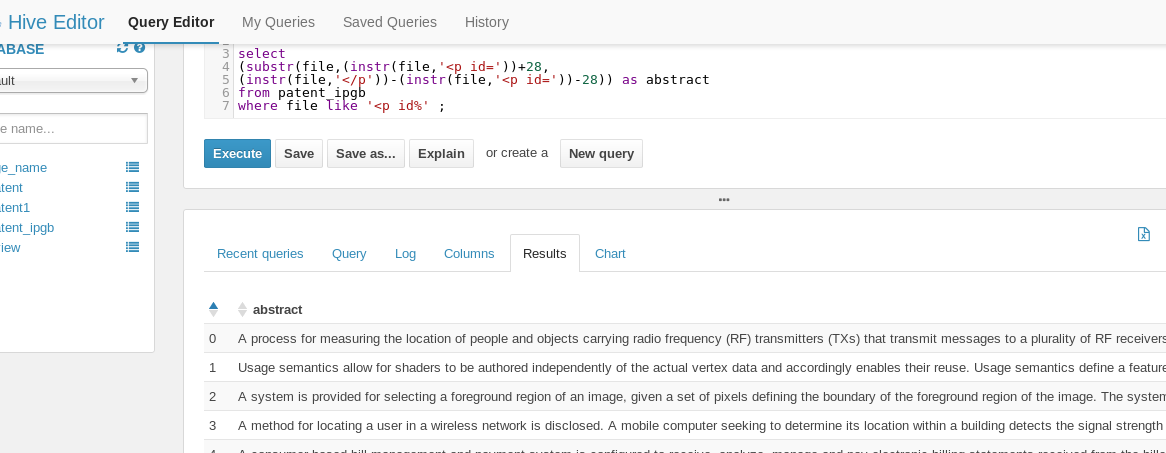
*(file,(instr(file,'<p id='))+28,*

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

*as abstract*

*frompatent\_ipgb*

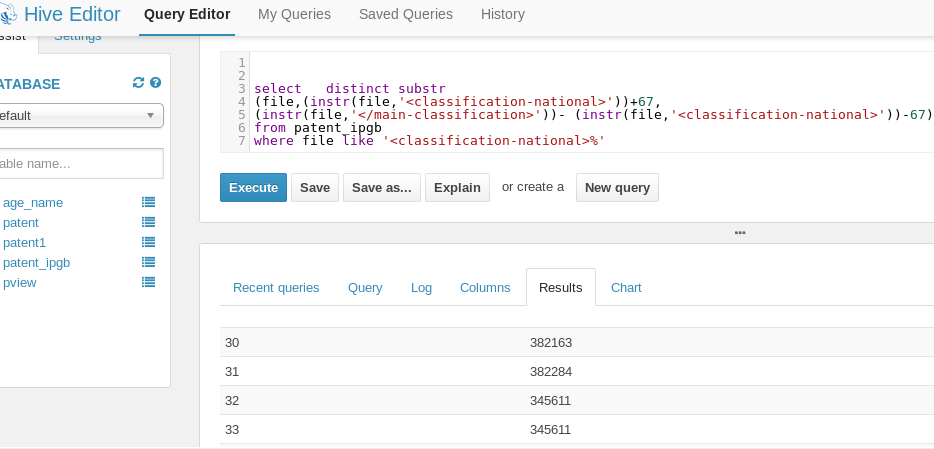
*where file like '<p id%' ;*

**

***Abstract fieldis 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>%';*

**

Code Appendix Phase-2

**Code used to derive the Factual Fields:**

**Code used to Extract Days Information:**

*select ((cast(substr(date\_grant,1,4) as int)-cast(substr(date\_publ,1,4) as int))\*365)*

*+*

*case when cast(substr(date\_grant,5,2) as int)>=cast(substr(date\_publ,5,2) as int)*

*then (cast(substr(date\_grant,5,2) as int)-cast(substr(date\_publ,5,2) as int))\*30*

*else (cast(substr(date\_publ,5,2) as int)-cast(substr(date\_grant,5,2) as int))\*30*

*end*

*+*

*case when cast(substr(date\_grant,7,2) as int)>=cast(substr(date\_publ,7,2) as int)*

*then (cast(substr(date\_grant,7,2) as int)-cast(substr(date\_publ,7,2) as int))*

*else (cast(substr(date\_publ,7,2) as int)-cast(substr(date\_grant,7,2) as int))*

*end as Granted\_days*

*from days*

-------------------------------------------------------------------------------------------------------------------------------

**Code used to Extract Backward and Forward Citations:**

*select substr(file,(instr(file,'le="'))+4,(instr(file,'.XML'))-(instr(file,'file='))-6) as patentno\_datepubl,*

*(substr(file,(instr(file,'<doc-number>'))+12,(instr(file,'</doc-number>'))-13)) as Backward\_citation,*

*case when (substr(file,(instr(file,'<category>'))+10,(instr(file,'</category>'))-11)) like 'cited by%'*

*then 1 else 0 end as Forward\_citation*

*from patent\_ipgb where file like '%us-patent-grant%' or file like '<category>%'*

-------------------------------------------------------------------------------------------------------------------------------

**Appendix-2**

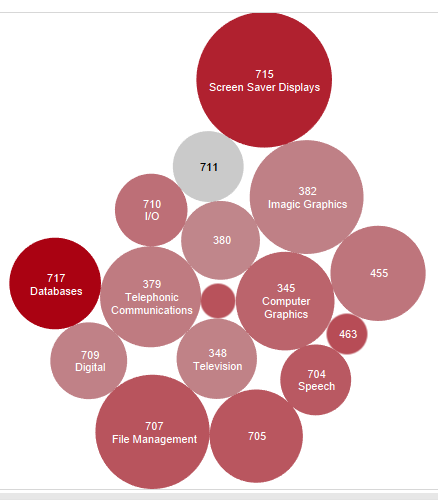
**Results supporting Strategy-1**

**Bubble Chart with Patent Count, Days to Granted and Class:**

We segregated patents based on their class and performed a bubble chart for each *Class* where *'Size of bubble indicates the Patent Count in that class'* and *'Thickness of color indicates the Average Days the patents in the class take to get granted'*

From the Visualization it is evident that patents under Class- Database took relatively more days to get granted. *Refer Fig-1*

**Fig-1**



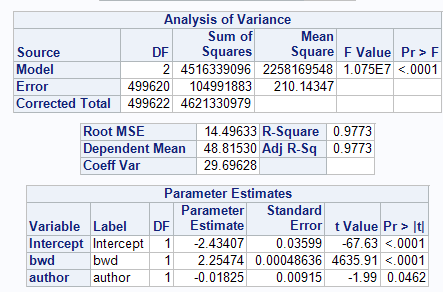
**Regression Results:**

**Regression:**

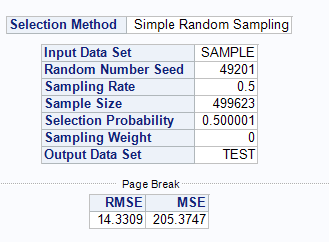
We wanted to make sure that whatever recommendations we make to improve the R&D in Class -Database does not impact the value of the patent. And apparently the patent is valued based on its Forward Citations.

So we ran a Regression Model to see what variables have impact on the Citations . We figured out that Backward Citation and No of Authors affected the Forward Citation count. But surprisingly for Microsoft the No of Authors/patent had a slight negative impact on the Forward Citation count. (*Refer Fig-2 and Fig-3* )

**Train results: Fig-2**



**Test Results: Fig-3**



**Regression Result Interpretation:**

* ***Model -> fwd=-2.4341+2.255\*bwd-0.01825\*author;***
* Difference between **Test** and **Train** data sets' MSE is just **0.95%.**
* Our Model Explains a variance of **97%.**
* All the variables in the model have passed the T-test α<0.05 and hence this overcomes the null hypothesis scenario. Since there at least one significant variable with β<>0.
* Our Model is significant since it has passed the F-test

**Recommendation based on Bubble Chart and Regression results:**

Based on the Model results we concluded that the resources should not be increased in the Class-Database R&D domain since it may affect the quality of the patents. Microsoft should instead allocate some funds from Class-Game Design(711) to Class-Database since from the Bubble Chart Viz we see that Class-Game Design has very Less patents and very Less days as well.

**Appendix-3**

**Strategy-2**

**Metrics for Top patents:**

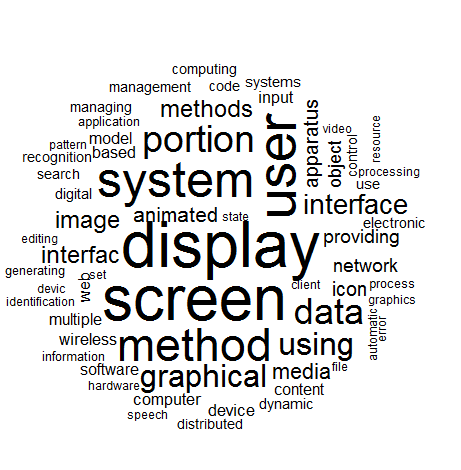
We derived our own metrics to figure out the top rated patents in Microsoft taking into consideration Forward and Backward Citations, Days to get Granted and No of Authors. *Forward Citations\*Backward Citations / No of days to get Granted\*No of Authors*.

**Word cloud:**

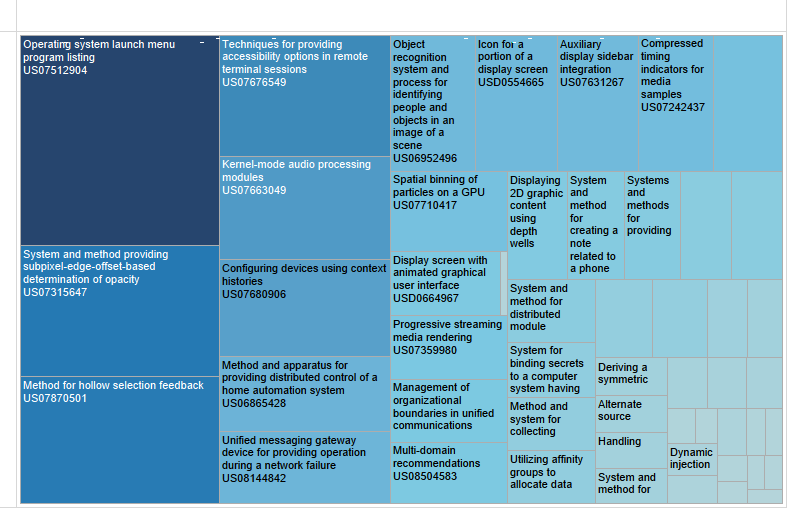
We ran word cloud on the data set which had the top Patents based on the above Metrics. We compared the Word cloud with the Word cloud ran on the complete data set. We found that both the Word clouds had the same Buzz words. Through this we concluded that Microsoft's Top patents are nothing but the areas where Microsoft is predominantly working on.

From the domains from the Word clouds ,we are just taking the *Operating System* and *Mobile* markets alone to compare and develop a strategy. *ReferFig-4 and Fig-6*

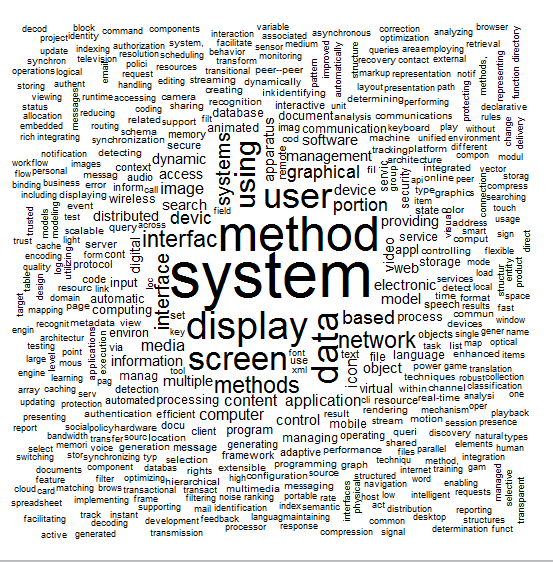
**Top patents World cloud of Microsoft Fig-4**



**Tree map for the top patents: Fig-5**



**Complete Microsoft Word cloud: Fig-6**

****

**Topic Modeling:**

We use results from Topic Modeling to support our argument. We consolidated the entire weights of each topic and have enlisted the top 10 topics based on their weights which has the same keywords as Word clouds. *Refer Table-1*

**Top Topics and their respective weights over all the Documents: Table-1**

|  |  |
| --- | --- |
| Total Weight in all docs | Topic/Words |
| 856.9399862 | system ,methods, computer, systems, storage, memory, file managing ,operating components |
| 881.5970101 | method ,system ,web document, data ,images, time ,providing, network text |
| 885.4349865 | data ,information, system, dynamic, peer ,resource structure, resources, execution creating |
| 919.4149812 | apparatus ,devices, image search, multi wireless ,process code, security detection |
| 819.2359869 | user interface ,device ,graphical software service, input ,animated, mobile, performance |
| 878.2960069 | display screen ,portion, icon ,coding, location, compression, secure, decoding, online |
| 851.2430036 | based ,management ,multiple application, object, database, state ,context results ,message |
| 882.472977 | video ,control access applications, method ,digital audio recognition, speech data |
| 909.2719819 | network system, electronic model language ,client services ,program protocol architecture |
| 881.3670216 | content ,media processing, objects environment, systems, computing, distributed communication providing |

**Insights based on our analysis:**

In short we are using different Technologies/methods to prove the same thing that *"Microsoft's Top patents are nothing but the areas where Microsoft is predominantly working on"*

From the data analysis done on the patent data we deduced that Microsoft is predominantly concentrated on Desktop/Laptop OS and Display screens.

Market share states that Windows are leading in the Desktop/Laptop OS market. But it is far behind in the Mobile market though we see an increase in Market share from 2.07 to 2.45%. Microsoft should invest a lot in Mobiles since we see potential growth. The first step they should take is to increase the number of Apps in the Windows Store. Because Smart Phones are all about the Apps and the Interface.

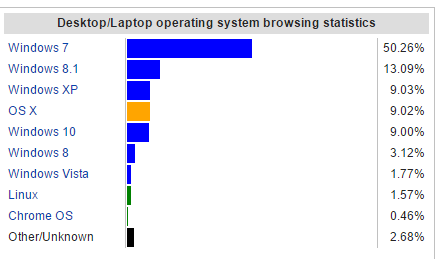
***.***

**Interpretation from the below figures:**

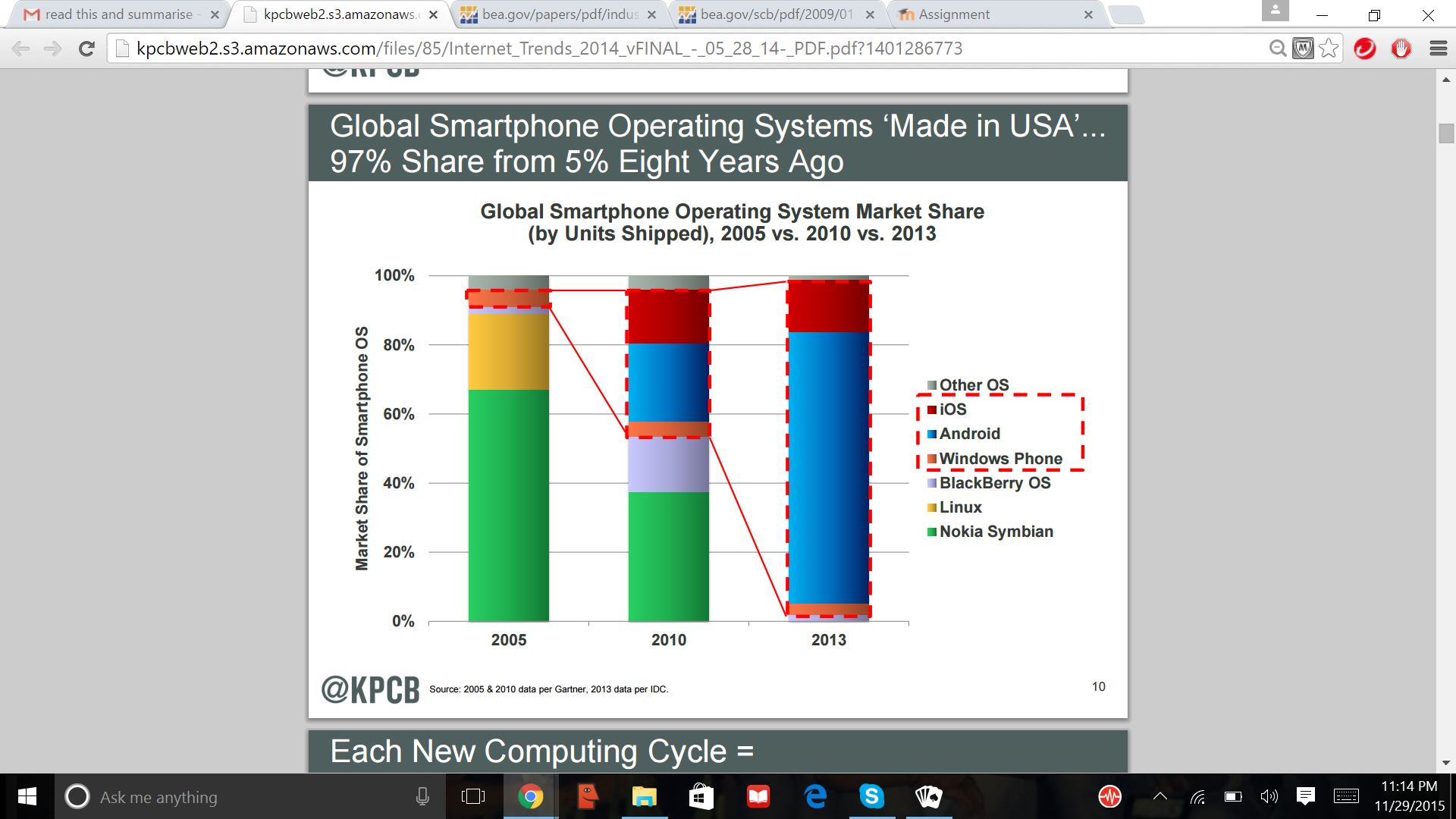
***The reason for the number of Apps being less in Windows Store is that there are not as many Developers who know C# than Java ,which is used to develop Android Apps***

**Market Data Analytics that supports our Argument:**

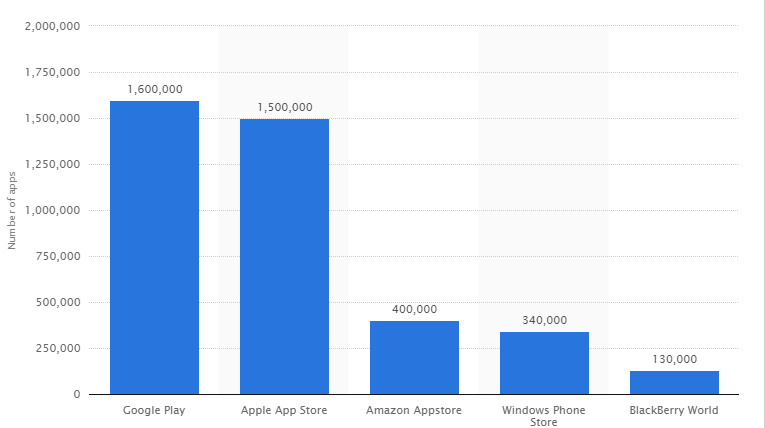
**OS Market: Fig-7**



**Mobile Market: Fig-8**

****

**Number of Apps in each App store: Fig-9**

****