



CSE408 –DATA MINING PROJECT  
ANALYZING ENTERPRISE DATA USING  
MAPREDUCE ON HADOOP

Project Report

Slot – F1

Faculty-

Dr. Selvakumar K

SCOPE

Submitted by –

14BCE0703 Pulkit Mittal

14BCE0733 Nikita Gill

## ABSTRACT

Big data is presently a buzzword throughout the software industry. Many IT giants like IBM, Google and Oracle have invested billions of dollars into the research to develop frameworks that can handle the big data efficiently. In this project, we make an attempt to analyze enterprise data that cannot be analyzed locally due to size and computation limitations. We analyze these datasets and try to extract the hidden insights from the same. The project aims to design an algorithm that can group and analyze the datasets as per the requirement. Based on the results of the program, we try to identify certain patterns with the help of data visualization and predict information.

## LITERATURE SURVEY

Early 2000s — Google stumbled across an obstacle while carrying out its mission — to organize information from all across the globe — which meant that it was crawling, copying, and indexing the entire Internet continuously. Back then, no software could handle the excessively large volume of data to be processed. Even Google's own custom infrastructure couldn't.

To deal with the situation, Google's engineers designed and built a new data processing infrastructure that comprised of two core components —the Google File System, or GFS, which provided fault-tolerant, reliable, and scalable storage, and MapReduce, a data processing system that allowed work to be split among large numbers of servers and carried out in parallel. Google published an academic paper [1] in 2004 describing its work.

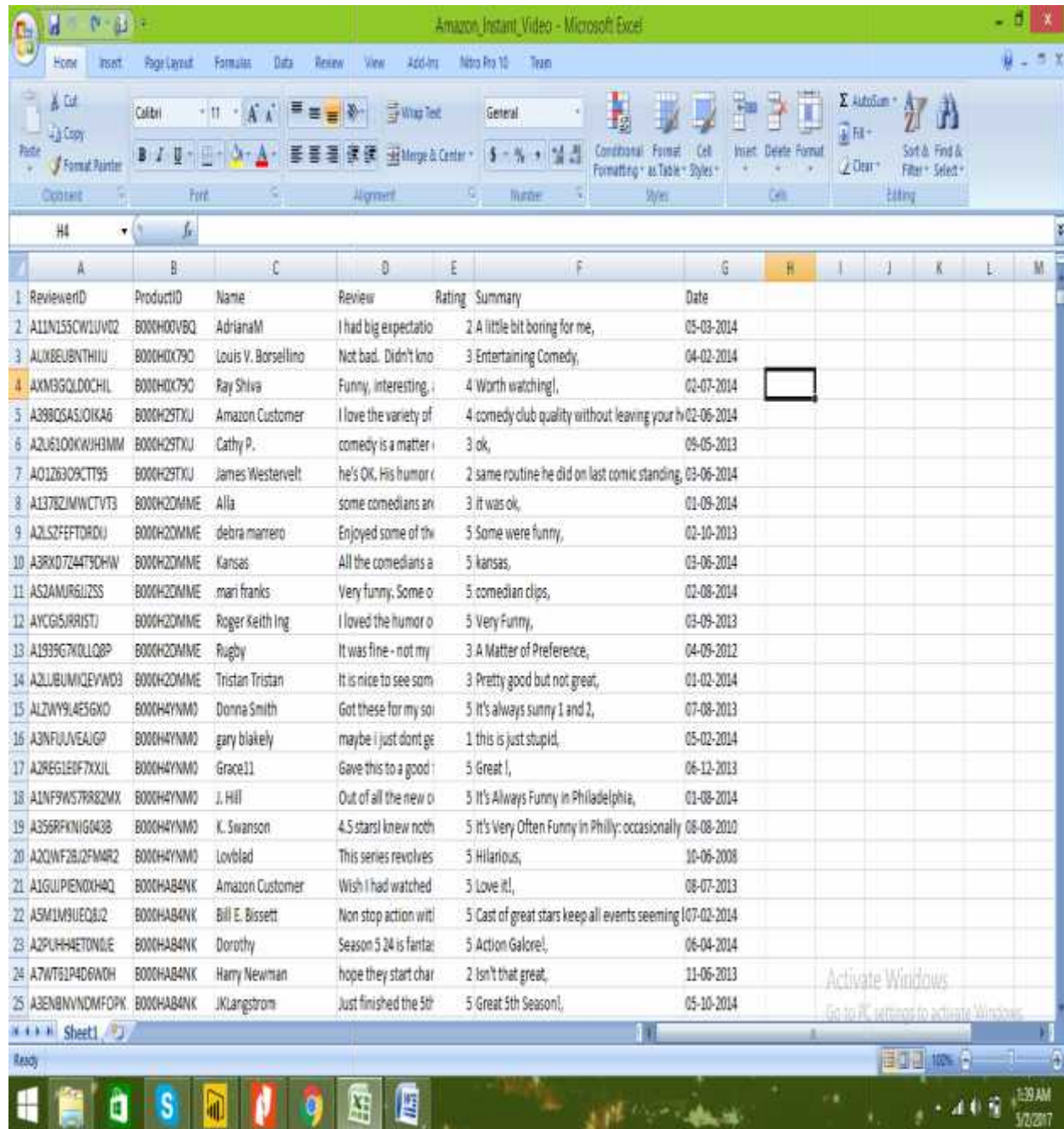
Doug Cutting, a well-known open source software developer, thereafter decided to use the technique Google's paper described. He was working on a web crawler called Nutch [2] and was having the same problems with data volumes and indexing speed that had driven Google to develop MapReduce. He replaced the data collection and processing infrastructure behind the crawler, basing his new implementation on MapReduce. He named the new software Hadoop, after a toy stuffed elephant that belonged to his young son.

Hadoop is an open source project [3] and operates under the Apache Software Foundation today. Hadoop has become a household name and is one of the most popular technologies today to handle big data. It is a data storage and analysis system which is scalable, incredibly flexible and works under the assumption that hardware failures are common occurrences and should be automatically handled by the framework [4] – an assumption that directly leads to its fault tolerant nature.

Hadoop can be deployed in a traditional on-site datacenter as well as in the cloud. Microsoft offers its cloud services via the Microsoft Azure Cloud Service platform which includes HDinsight – the service which shall be used in this project to create and deploy clusters as well as run

mapreduce jobs on the data that needs to be analyzed. HDinsight offers efficient, reliable and performance centric results [5] with a pay-per-use model which is perfect for a project like the one we are aiming for.

## Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M
	ReviewerID	ProductID	Name	Review	Rating	Summary	Date						
1	A11N155CWJUV02	B000H00VBQ	AdrianaM	I had big expectatio	2	A little bit boring for me,	05-03-2014						
2	ALX8EUBNTHIU	B000H0X79C	Louis V. Borsellino	Not bad. Didn't kno	3	Entertaining Comedy,	04-02-2014						
3	AXM9GQLD0CHL	B000H0X79C	Ray Shiva	Funny, interesting, i	4	Worth watching!.	02-07-2014						
4	A398QASJ0IKA6	B000H29TXU	Amazon Customer	I love the variety of	4	comedy club quality without leaving your h	02-06-2014						
5	A2U6100KJUH3NM	B000H29TXU	Cathy P.	comedy is a matter	3	ok,	09-05-2013						
6	A0126309CTT95	B000H29TXU	James Westervelt	he's OK. His humor	2	same routine he did on last comic standing,	03-06-2014						
7	A13782IMNCTVT3	B000H20MME	Alla	some comedians an	3	it was ok,	01-09-2014						
8	A2LSZFEFTORDIU	B000H20MME	debra marrero	Enjoyed some of th	5	Some were funny,	02-10-2013						
9	A3RXD7Z4479CHW	B000H20MME	Kansas	All the comedians a	5	kansas,	03-06-2014						
10	AS2AMUR6J2SS	B000H20MME	man franks	Very funny. Some o	5	comedian clips,	02-08-2014						
11	AYCG5JRRISTJ	B000H20MME	Roger Keith Ing	I loved the humor o	5	Very Funny,	03-09-2013						
12	A193G7K0LQBP	B000H20MME	Rugby	It was fine - not my	3	A Matter of Preference,	04-09-2012						
13	A2LUBUMQEVWD3	B000H20MME	Tristan Tristan	It is nice to see som	3	Pretty good but not great,	01-02-2014						
14	ALZWY9L4E5GXO	B000H4YNM0	Donna Smith	Got these for my soi	5	It's always sunny 1 and 2,	07-08-2013						
15	A3NFUJVEAJGP	B000H4YNM0	gary blakely	maybe i just dont ge	1	this is just stupid,	05-02-2014						
16	A2REG1E0F7XJL	B000H4YNM0	Grace11	Gave this to a good	5	Great I,	06-12-2013						
17	A1NF9W57R82MX	B000H4YNM0	J. Hill	Out of all the new o	5	It's Always Funny in Philadelphia,	01-08-2014						
18	A356RFXNIG0438	B000H4YNM0	K. Swanson	4.5 stars! knew noth	5	It's Very Often Funny in Philly; occasionally	08-08-2010						
19	A2QWF2B2FMMR2	B000H4YNM0	Loveblad	This series revolves	5	Hilarious,	10-06-2008						
20	A1GUJPIEN0XHAQ	B000HAB4NK	Amazon Customer	Wish I had watched	3	Love it!,	08-07-2013						
21	ASM1M9UEQ8J2	B000HAB4NK	Bill E. Bissett	Non stop action witi	5	Cast of great stars keep all events seeming	10-02-2014						
22	A2PUH4HETN0UE	B000HAB4NK	Dorothy	Season 5 24 is fantas	5	Action Galore!,	06-04-2014						
23	A7W761P4D6W0H	B000HAB4NK	Harry Newman	hope they start char	2	Isn't that great,	11-06-2013						
24	A3ENBNVNDMF0PK	B000HAB4NK	JKLangstrom	Just finished the 5th	5	Great 5th Season!,	05-10-2014						

## HARDWARE AND SOFTWARE REQUIREMENTS

Since this project is dependent on cloud services for the analysis, requirements are:

**Hardware Requirements** – Any system that can run Microsoft's Azure services and have enough storage space to handle the data being analyzed can be used.

**Software Requirements** – Microsoft Azure's HDInsight platform for Cluster creation and Deployment, Apache Hadoop for analyzing the data and handling data using HDFS, Tools such as Microsoft Power BI for data visualization.

## SOFTWARE ARCHITECTURE AND DESIGN

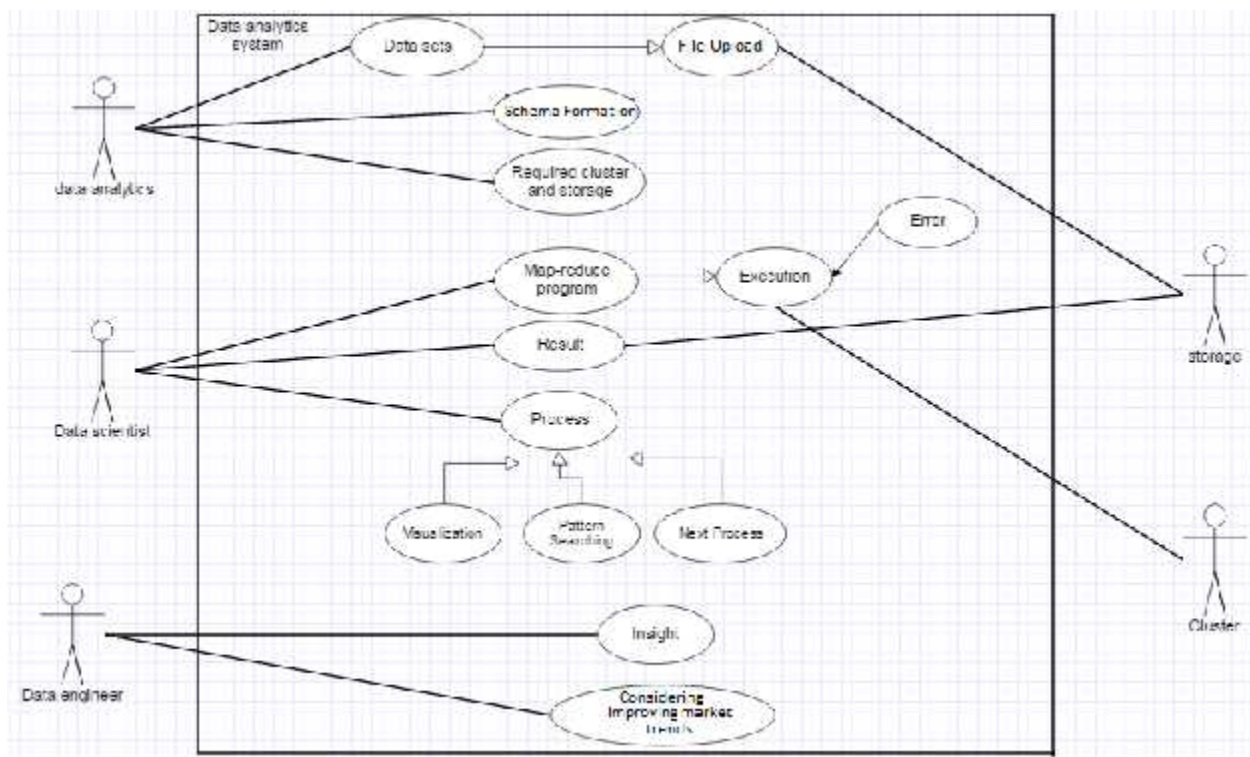


Figure 1. A Use-Case Diagram that illustrates the actors and various use-cases.

This project uses Microsoft's Azure Cloud Service which provides us with a multitude of services. HDInsight, specifically, is the service that is being used in this project.

Azure HDInsight is a service that deploys Hadoop on Microsoft Azure. HDInsight uses Hortonworks HDP and was jointly developed for HDI with Hortonworks. HDInsight also supports creation of Hadoop clusters using Linux with Ubuntu. HDInsight is very flexible in its usage and we can, at any point of time, scale up the cluster by increasing the

amount or type of worker/head nodes that exist in the cluster.

HDInsight is used to set up and deploy clusters on cloud. These clusters are comprised of head as well as worker nodes. Once the cluster is set up along with the storage container, the cluster can be used to upload/download data as well as to run programs that, on a fundamental level, utilize mapreduce to perform analytical tasks.

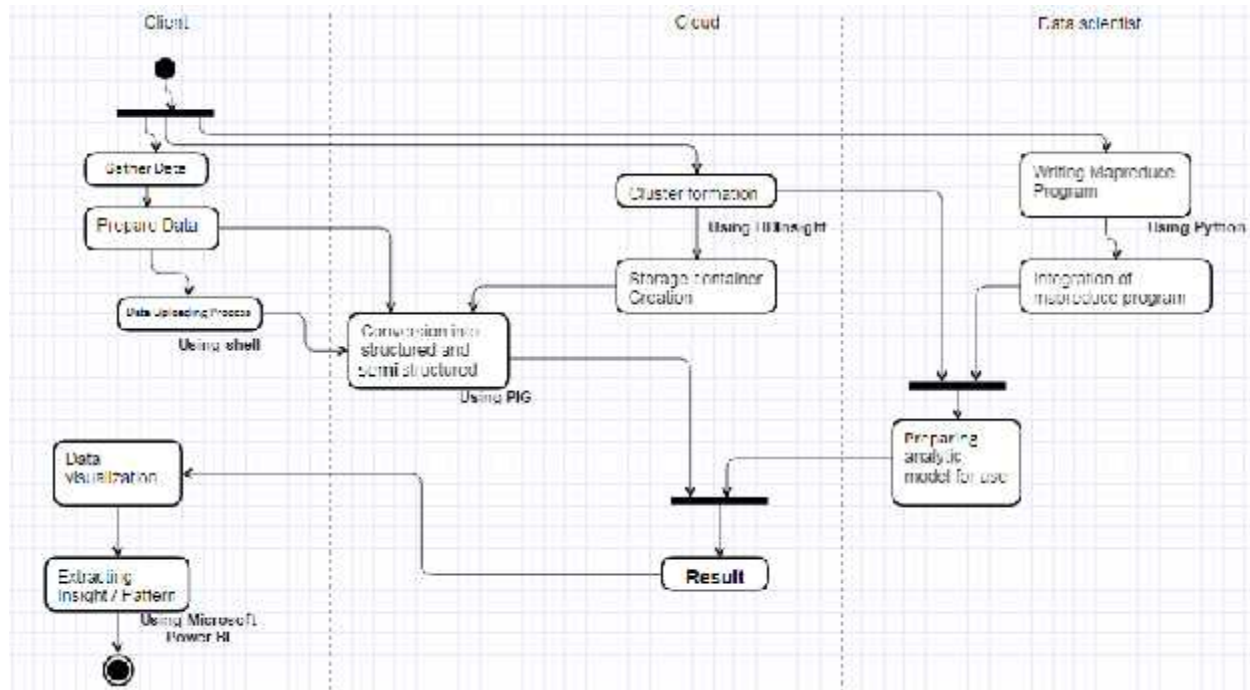


Figure 2. Activity Diagram that covers the entire workflow – from setting up a cloud-based cluster, to managing data on the storage, its analysis and visualization of the results obtained from analyzing it.

As shown in Figure 2, the initial step comprises of gathering data that needs to be analyzed. Depending on the size of the data and scope of the analytical process which determines the processing power needed, a cluster is deployed using HDInsight on the cloud. The Hadoop cluster, depending on the user's need, can be deployed as a cluster utilizing Linux-based machines or Windows-based machines. And depending on the choice made, HDInsight offers different services unique to the selected OS. For example, HDInsight offers SSH access to the Linux clusters while the Windows clusters have exclusive access to the Remote Desktop Access functionality which is otherwise missing from the Linux-based clusters. Since Microsoft's toolset for a Big Data Analyst comprises of a variety of tools, the entire process of cluster creation and deployment can be automated using a script by utilizing Microsoft's Powershell service. The underlying Hadoop framework handles distribution of data for storage using its HDFS

and distribution of analytical work with MapReduce framework.

Once the cluster has been deployed, data to be analyzed is uploaded to the cluster via Azure Command Line Interface. This data is then analyzed by running a mapreduce program on the various machines present on the cluster. The result is then either analyzed further for more efficiency or downloaded locally for it to be used for data visualization. Power BI is a free tool that is then used on the output of the analysis to obtain the results in a graphical format ie, pie-chart, bar graph etc.

## MODULE DESCRIPTION

The modules comprise of the following aspects –

ANALYSIS

DATA PROCESSING

VISUALIZATION

They are discussed below:

Data Processing (using MapReduce)

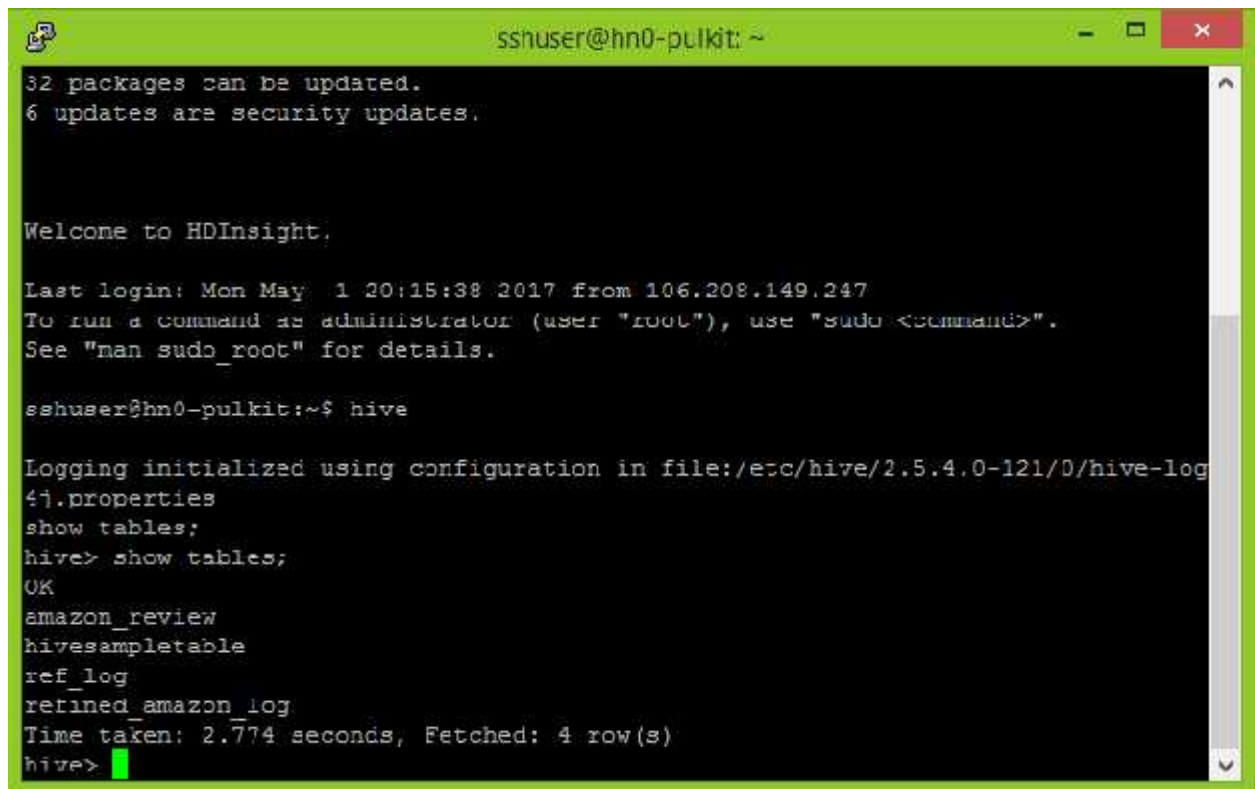
This process forms the core of data analysis module. It includes both the MapReduce programs written in order to be used with the Tez engine that perform the Map & Reduce jobs when analyzing data using pig scripts or hive queries as well as the mapreduce programs that can be used to perform specific tasks given structured data.

The former forms a part of the package that our environment provides us along with Hive/Pig. The Hive/Pig scripts are run and processing is carried out by carrying out Map and Reduce jobs with the help of these programs. At no point of time does the data analyst explicitly evokes these programs.

The latter includes programs written explicitly in java like wordcount that form a part of specific mapreduce programs that the package provides analysts. These programs have to be explicitly evoked and data passed to them for them to process it and return useful data.



```
1 001 1
2 002 1
3 003 1
4 004 1
5 005 1
6 006 1
7 007 1
8 008 1
9 009 1
10 010 1
11 011 1
12 012 1
13 013 1
14 014 1
15 015 1
16 016 1
17 017 1
18 018 1
19 019 1
20 020 1
21 021 1
22 022 1
23 023 1
24 024 1
25 025 1
26 026 1
27 027 1
28 028 1
29 029 1
30 030 1
31 031 1
32 032 1
33 033 1
34 034 1
35 035 1
36 036 1
37 037 1
38 038 1
39 039 1
40 040 1
41 041 1
42 042 1
43 043 1
44 044 1
45 045 1
46 046 1
47 047 1
48 048 1
49 049 1
50 050 1
51 051 1
52 052 1
53 053 1
54 054 1
55 055 1
56 056 1
57 057 1
58 058 1
59 059 1
60 060 1
61 061 1
62 062 1
63 063 1
64 064 1
65 065 1
66 066 1
67 067 1
68 068 1
69 069 1
70 070 1
71 071 1
72 072 1
73 073 1
74 074 1
75 075 1
76 076 1
77 077 1
78 078 1
79 079 1
80 080 1
81 081 1
82 082 1
83 083 1
84 084 1
85 085 1
86 086 1
87 087 1
88 088 1
89 089 1
90 090 1
91 091 1
92 092 1
93 093 1
94 094 1
95 095 1
96 096 1
97 097 1
98 098 1
99 099 1
100 100 1
101 101 1
102 102 1
103 103 1
104 104 1
105 105 1
106 106 1
107 107 1
108 108 1
109 109 1
110 110 1
111 111 1
112 112 1
113 113 1
114 114 1
115 115 1
116 116 1
117 117 1
118 118 1
119 119 1
120 120 1
121 121 1
122 122 1
123 123 1
124 124 1
125 125 1
126 126 1
127 127 1
128 128 1
129 129 1
130 130 1
131 131 1
132 132 1
133 133 1
134 134 1
135 135 1
136 136 1
137 137 1
138 138 1
139 139 1
140 140 1
141 141 1
142 142 1
143 143 1
144 144 1
145 145 1
146 146 1
147 147 1
148 148 1
149 149 1
150 150 1
151 151 1
152 152 1
153 153 1
154 154 1
155 155 1
156 156 1
157 157 1
158 158 1
159 159 1
160 160 1
161 161 1
162 162 1
163 163 1
164 164 1
165 165 1
166 166 1
167 167 1
168 168 1
169 169 1
170 170 1
171 171 1
172 172 1
173 173 1
174 174 1
175 175 1
176 176 1
177 177 1
178 178 1
179 179 1
180 180 1
181 181 1
182 182 1
183 183 1
184 184 1
185 185 1
186 186 1
187 187 1
188 188 1
189 189 1
190 190 1
191 191 1
192 192 1
193 193 1
194 194 1
195 195 1
196 196 1
197 197 1
198 198 1
199 199 1
200 200 1
201 201 1
202 202 1
203 203 1
204 204 1
205 205 1
206 206 1
207 207 1
208 208 1
209 209 1
210 210 1
211 211 1
212 212 1
213 213 1
214 214 1
215 215 1
216 216 1
217 217 1
218 218 1
219 219 1
220 220 1
221 221 1
222 222 1
223 223 1
224 224 1
225 225 1
226 226 1
227 227 1
228 228 1
229 229 1
230 230 1
231 231 1
232 232 1
233 233 1
234 234 1
235 235 1
236 236 1
237 237 1
238 238 1
239 239 1
240 240 1
241 241 1
242 242 1
243 243 1
244 244 1
245 245 1
246 246 1
247 247 1
248 248 1
249 249 1
250 250 1
251 251 1
252 252 1
253 253 1
254 254 1
255 255 1
256 256 1
257 257 1
258 258 1
259 259 1
260 260 1
261 261 1
262 262 1
263 263 1
264 264 1
265 265 1
266 266 1
267 267 1
268 268 1
269 269 1
270 270 1
271 271 1
272 272 1
273 273 1
274 274 1
275 275 1
276 276 1
277 277 1
278 278 1
279 279 1
280 280 1
281 281 1
282 282 1
283 283 1
284 284 1
285 285 1
286 286 1
287 287 1
288 288 1
289 289 1
290 290 1
291 291 1
292 292 1
293 293 1
294 294 1
295 295 1
296 296 1
297 297 1
298 298 1
299 299 1
300 300 1
301 301 1
302 302 1
303 303 1
304 304 1
305 305 1
306 306 1
307 307 1
308 308 1
309 309 1
310 310 1
311 311 1
312 312 1
313 313 1
314 314 1
315 315 1
316 316 1
317 317 1
318 318 1
319 319 1
320 320 1
321 321 1
322 322 1
323 323 1
324 324 1
325 325 1
326 326 1
327 327 1
328 328 1
329 329 1
330 330 1
331 331 1
332 332 1
333 333 1
334 334 1
335 335 1
336 336 1
337 337 1
338 338 1
339 339 1
340 340 1
341 341 1
342 342 1
343 343 1
344 344 1
345 345 1
346 346 1
347 347 1
348 348 1
349 349 1
350 350 1
351 351 1
352 352 1
353 353 1
354 354 1
355 355 1
356 356 1
357 357 1
358 358 1
359 359 1
360 360 1
361 361 1
362 362 1
363 363 1
364 364 1
365 365 1
366 366 1
367 367 1
368 368 1
369 369 1
370 370 1
371 371 1
372 372 1
373 373 1
374 374 1
375 375 1
376 376 1
377 377 1
378 378 1
379 379 1
380 380 1
381 381 1
382 382 1
383 383 1
384 384 1
385 385 1
386 386 1
387 387 1
388 388 1
389 389 1
390 390 1
391 391 1
392 392 1
393 393 1
394 394 1
395 395 1
396 396 1
397 397 1
398 398 1
399 399 1
400 400 1
401 401 1
402 402 1
403 403 1
404 404 1
405 405 1
406 406 1
407 407 1
408 408 1
409 409 1
410 410 1
411 411 1
412 412 1
413 413 1
414 414 1
415 415 1
416 416 1
417 417 1
418 418 1
419 419 1
420 420 1
421 421 1
422 422 1
423 423 1
424 424 1
425 425 1
426 426 1
427 427 1
428 428 1
429 429 1
430 430 1
431 431 1
432 432 1
433 433 1
434 434 1
435 435 1
436 436 1
437 437 1
438 438 1
439 439 1
440 440 1
441 441 1
442 442 1
443 443 1
444 444 1
445 445 1
446 446 1
447 447 1
448 448 1
449 449 1
450 450 1
451 451 1
452 452 1
453 453 1
454 454 1
455 455 1
456 456 1
457 457 1
458 458 1
459 459 1
460 460 1
461 461 1
462 462 1
463 463 1
464 464 1
465 465 1
466 466 1
467 467 1
468 468 1
469 469 1
470 470 1
471 471 1
472 472 1
473 473 1
474 474 1
475 475 1
476 476 1
477 477 1
478 478 1
479 479 1
480 480 1
481 481 1
482 482 1
483 483 1
484 484 1
485 485 1
486 486 1
487 487 1
488 488 1
489 489 1
490 490 1
491 491 1
492 492 1
493 493 1
494 494 1
495 495 1
496 496 1
497 497 1
498 498 1
499 499 1
500 500 1
501 501 1
502 502 1
503 503 1
504 504 1
505 505 1
506 506 1
507 507 1
508 508 1
509 509 1
510 510 1
511 511 1
512 512 1
513 513 1
514 514 1
515 515 1
516 516 1
517 517 1
518 518 1
519 519 1
520 520 1
521 521 1
522 522 1
523 523 1
524 524 1
525 525 1
526 526 1
527 527 1
528 528 1
529 529 1
530 530 1
531 531 1
532 532 1
533 533 1
534 534 1
535 535 1
536 536 1
537 537 1
538 538 1
539 539 1
540 540 1
541 541 1
542 542 1
543 543 1
544 544 1
545 545 1
546 546 1
547 547 1
548 548 1
549 549 1
550 550 1
551 551 1
552 552 1
553 553 1
554 554 1
555 555 1
556 556 1
557 557 1
558 558 1
559 559 1
560 560 1
561 561 1
562 562 1
563 563 1
564 564 1
565 565 1
566 566 1
567 567 1
568 568 1
569 569 1
570 570 1
571 571 1
572 572 1
573 573 1
574 574 1
575 575 1
576 576 1
577 577 1
578 578 1
579 579 1
580 580 1
581 581 1
582 582 1
583 583 1
584 584 1
585 585 1
586 586 1
587 587 1
588 588 1
589 589 1
590 590 1
591 591 1
592 592 1
593 593 1
594 594 1
595 595 1
596 596 1
597 597 1
598 598 1
599 599 1
600 600 1
601 601 1
602 602 1
603 603 1
604 604 1
605 605 1
606 606 1
607 607 1
608 608 1
609 609 1
610 610 1
611 611 1
612 612 1
613 613 1
614 614 1
615 615 1
616 616 1
617 617 1
618 618 1
619 619 1
620 620 1
621 621 1
622 622 1
623 623 1
624 624 1
625 625 1
626 626 1
627 627 1
628 628 1
629 629 1
630 630 1
631 631 1
632 632 1
633 633 1
634 634 1
635 635 1
636 636 1
637 637 1
638 638 1
639 639 1
640 640 1
641 641 1
642 642 1
643 643 1
644 644 1
645 645 1
646 646 1
647 647 1
648 648 1
649 649 1
650 650 1
651 651 1
652 652 1
653 653 1
654 654 1
655 655 1
656 656 1
657 657 1
658 658 1
659 659 1
660 660 1
661 661 1
662 662 1
663 663 1
664 664 1
665 665 1
666 666 1
667 667 1
668 668 1
669 669 1
670 670 1
671 671 1
672 672 1
673 673 1
674 674 1
675 675 1
676 676 1
677 677 1
678 678 1
679 679 1
680 680 1
681 681 1
682 682 1
683 683 1
684 684 1
685 685 1
686 686 1
687 687 1
688 688 1
689 689 1
690 690 1
691 691 1
692 692 1
693 693 1
694 694 1
695 695 1
696 696 1
697 697 1
698 698 1
699 699 1
700 700 1
701 701 1
702 702 1
703 703 1
704 704 1
705 705 1
706 706 1
707 707 1
708 708 1
709 709 1
710 710 1
711 711 1
712 712 1
713 713 1
714 714 1
715 715 1
716 716 1
717 717 1
718 718 1
719 719 1
720 720 1
721 721 1
722 722 1
723 723 1
724 724 1
725 725 1
726 726 1
727 727 1
728 728 1
729 729 1
730 730 1
731 731 1
732 732 1
733 733 1
734
```

A terminal window with a green title bar and a black background. The window title is 'sshuser@hn0-pulkit: ~'. The text in the terminal is as follows:

```
32 packages can be updated.
6 updates are security updates.

Welcome to HDInsight.

Last login: Mon May 1 20:15:38 2017 from 106.208.149.247
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

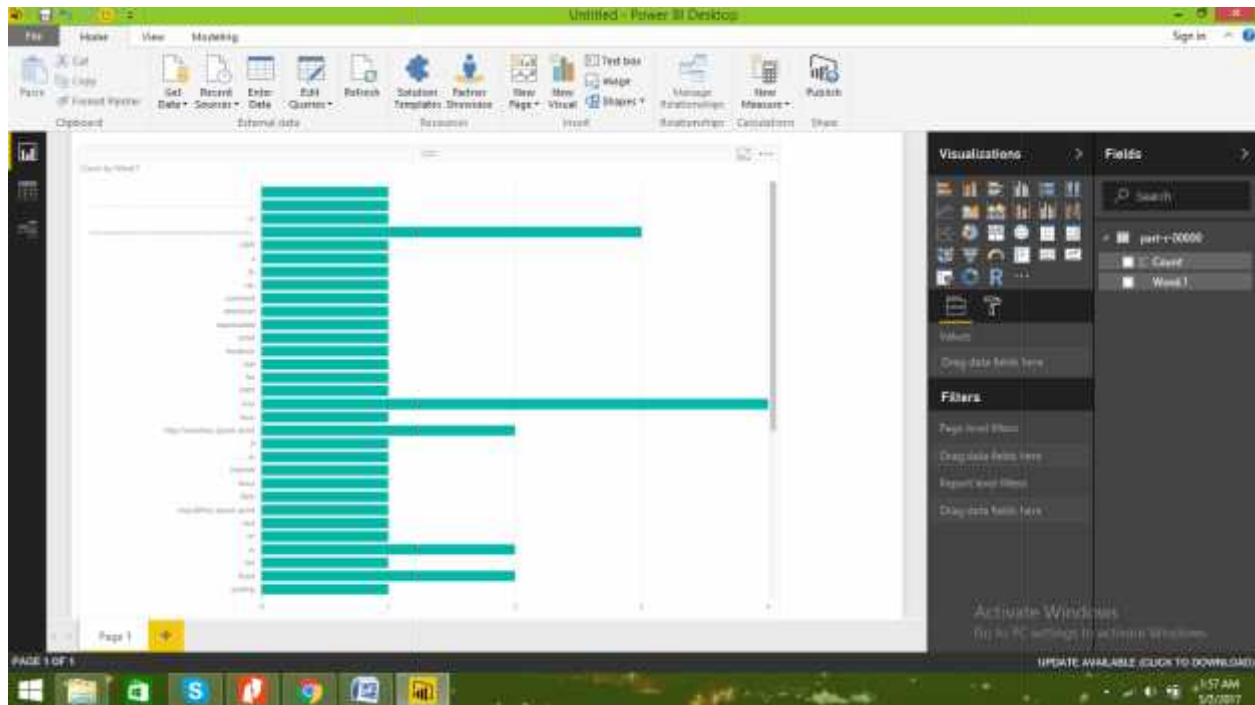
sshuser@hn0-pulkit:~$ hive

Logging initialized using configuration in file:/etc/hive/2.5.4.0-121.0/hive-log
41.properties
show tables;
hive> show tables;
OK
amazon_review
hivesampletable
ref_log
refined_amazon_log
Time taken: 2.774 seconds, Fetched: 4 row(s)
hive>
```

## Visualization (using PowerBi)

PowerBi is a data visualization tool that is provided free of cost by Microsoft. It can be used to visualize structured data such as text files with tab delimited data or data in csv formats. It is a very powerful tool that gives the user a lot of freedom when it comes to the kind of data visualization that can be used. It also enables drilling down and rolling up operations that can be used to abstract data or get into the details of the visualized data. As a result, PowerBi is the software of choice for the data visualization aspect of our project.





## SAMPLE CODE

The hive script below has been used to convert an amazon dataset that was originally in the json encoded format into a tab delimited structured format that can now be easily used for data visualization in PowerBi. It is basically performing the task of dataset refinement. Apart from that, it also uses this staged data to perform the task of counting the total number of 1 – 5 star ratings present in the video reviews dataset.

### Hive Script(s) –

```
CREATE EXTERNAL TABLE Amazon_Review
```

```
(ReviewerID STRING,
```

```
ProductID STRING,
```

```
Name STRING,
```

```
Review STRING,
```

```
Rating INT,
```

```
Summary STRING,
```

```
Period STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE LOCATION '/data/Amazon_Review';
```

```
LOAD DATA INPATH '/data/Amazon_Instant_Video.txt' INTO TABLE Amazon_Review;
```

```
CREATE TABLE Refined_Amazon_Log
(ReviewerID STRING,
ProductID STRING,
Name STRING,
Review STRING,
Rating INT,
Summary STRING,
Period STRING);
```

```
INSERT INTO TABLE Refined_Amazon_Log
SELECT * FROM Amazon_Review
WHERE Rating IS NOT NULL;
```

```
CREATE VIEW Ref_Log
AS
SELECT from_unixtime(unix_timestamp(Period, 'dd/MM/yyyy hh:mm:ss')) AS Period, Review,
Rating, Summary
FROM Refined_Amazon_Log;
```

```
SELECT CAST(SUBSTR(Period, 1, 10) AS date) AS Period, Rating, COUNT(*) AS Rating
FROM Ref_Log
GROUP BY CAST(SUBSTR(Period, 1, 10) AS date), Rating
ORDER BY Period, Rating;
```

```
sshuser@hnc-pulkit: ~  
METHOD                                DURATION(ms)  
parse                                11  
semanticAnalyze                      2,542  
TezBuildDag                          1,479  
TezSubmitToRunningDag               207  
TotalPrepTime                        5,933  
  
VIRIICES          TOTAL_TASKS  FAILED_ATTEMPTS  KILLED_TASKS  DURATION_SECONDS  C  
PU_TIME_MILLIS    GC_TIME_MILLIS    INPUT_RECORDS    OUTPUT_RECORDS  
Map 1             1             0             0             1.92  
3,600             118             1,000          5  
Reducer 2         1             0             0             0.63  
1,290             0             5             5  
Reducer 3         1             0             0             1.25  
1,290             145             5             0  
OK  
NULL 1 22  
NULL 2 23  
NULL 3 92  
NULL 4 204  
NULL 5 659  
Time taken: 11.493 seconds, Fetched: 5 row(s)  
hive>
```

## References:

1. Dean, J. and Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters." Appeared in Proceedings of the Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004. Available online at <http://labs.google.com/papers/mapreduce.html>, March 2010
2. Apache Nutch project, <http://lucene.apache.org/nutch/>, March 2010
3. Apache Hadoop project, <http://hadoop.apache.org/>, March 2010.
4. Mike Olson, "HADOOP: Scalable, Flexible Data Storage and Analysis". Appeared in IQT Quarterly, Spring 2010. Available Online at [https://blog.cloudera.com/wp-content/uploads/2010/05/Olson\\_IQT\\_Quarterly\\_Spring\\_2010.pdf](https://blog.cloudera.com/wp-content/uploads/2010/05/Olson_IQT_Quarterly_Spring_2010.pdf)
5. Bhardwaj, Aditya et al, "Analyzing BigData with Hadoop cluster in HDInsight azure Cloud". Appeared in 2015 Annual IEEE India Conference (INDICON), December 2015