ANALYSIS ON H1B VISA APPLICATIONS

Using Hadoop Ecosystem

Presented By:

Shannon Michelle D'souza

Student ID : S180010900133

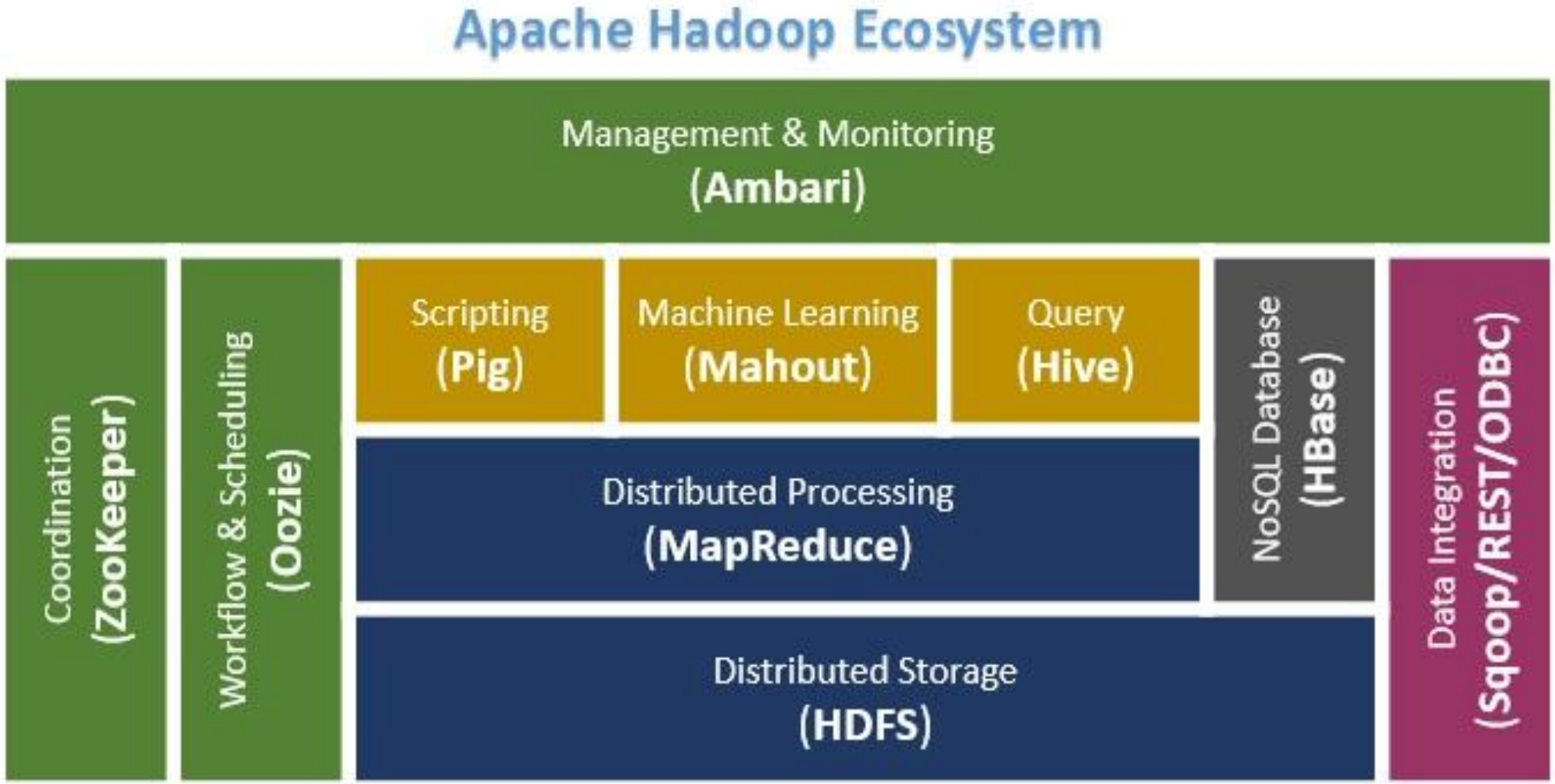
Registration No : R180010900252

NIIT Centre : Borivali West

Apache Hadoop

- Open source
- Java-based programming framework
- Distributed environment
- Parallel processing

Hadoop Ecosystem



Project Description

- H1B: an employment-based, non-immigrant visa category for temporary foreign workers in the United States.
- Applied by the Employer
- Commonly applied by International students
- Data set : 3 million records
- Time period : 2011 - 2016

Project Demonstration

- Query 1_A
- Query 1_B
- Query 2_A
- Query 2_B
- Query 3
- Query 4
- Query 5
- Query 6
- Query 7
- Query 8
- Query 9
- Query 10
- Query 11

MapReduce

Hive

Pig

Sqoop

Query 1_A

OBJECTIVE: To find if the number of petitions with Data Engineer job title is increasing over time

TECHNOLOGY USED: MapReduce (in Java)

EXPECTED OUTPUT FORMAT:

<year> <application count> <growth percentage>

Query 1_B

OBJECTIVE: To find the top 5 job titles which have the highest average growth in applications

TECHNOLOGY USED: MapReduce (in Java)

EXPECTED OUTPUT FORMAT:

<job title> <average growth percentage>

Query 2_A

OBJECTIVE: To find which part of the US has the most Data Engineer jobs for each year

TECHNOLOGY USED: Hive

EXPECTED OUTPUT FORMAT:

<worksite> <year> <application count>

Query 2_B

OBJECTIVE: To find the top 5 locations in the US which have got certified visa for each year

TECHNOLOGY USED: Pig

EXPECTED OUTPUT FORMAT:

<worksite> <case status> <year> <application count>

Query 3

OBJECTIVE: To find which industry (SOC Name) has the most number of Data Scientist positions

TECHNOLOGY USED: Hive

EXPECTED OUTPUT FORMAT:

<soc name> <application count>

Query 4

OBJECTIVE: To find the top 5 employers that file the most number of petitions each year

TECHNOLOGY USED: Pig

EXPECTED OUTPUT FORMAT:

<year> <employer name> <application count>

Query 5

OBJECTIVE: To find the most popular top 10 job positions for H1B visa applications for each year

a) for all the applications

b) for only certified applications

TECHNOLOGY USED: Hive

EXPECTED OUTPUT FORMAT:

<year> <job title> <application count>

Query 6

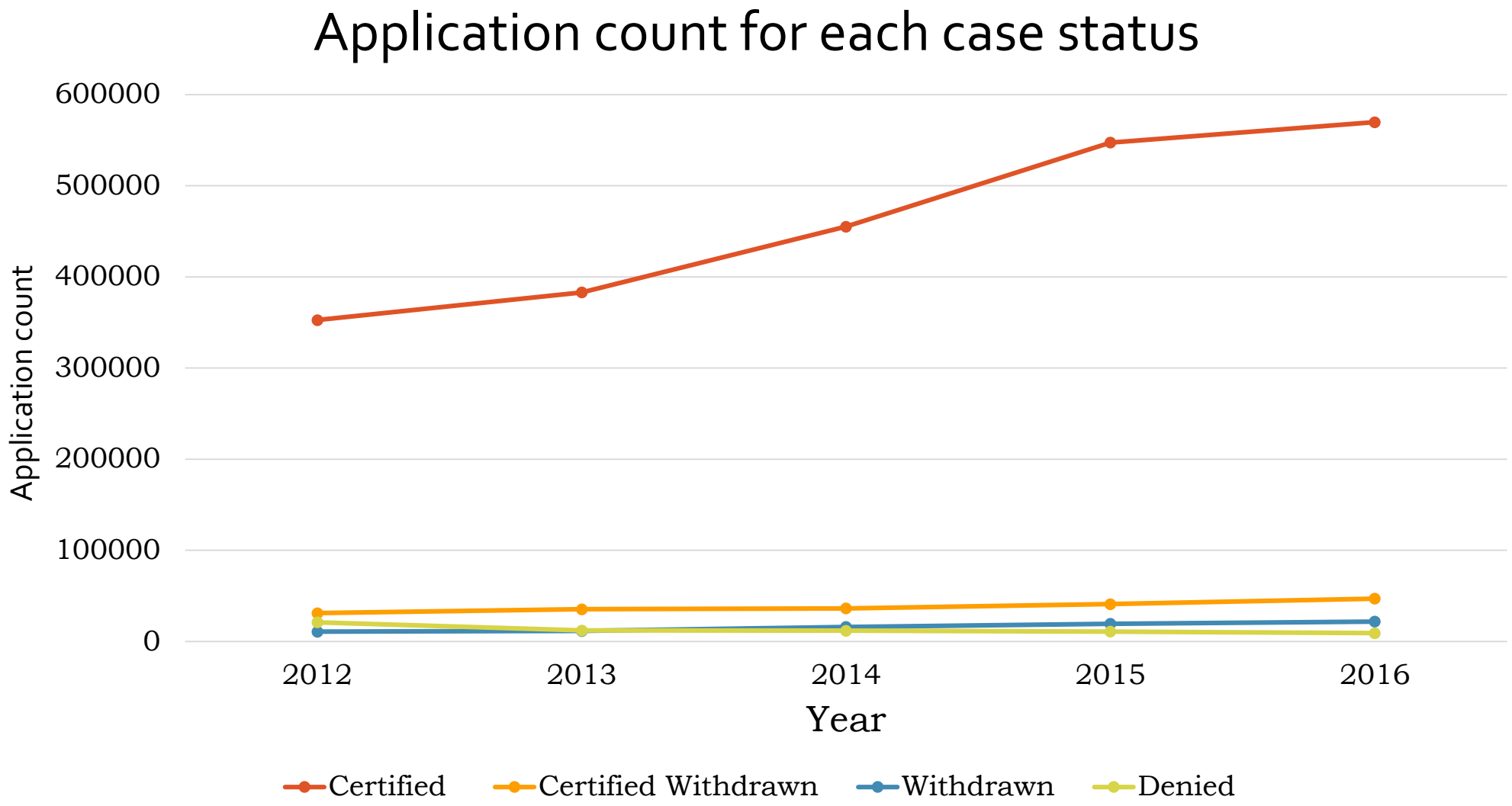
OBJECTIVE: To find the count and the percentage of each case status on total applications for each year. Also, to create a line graph depicting the pattern of ALL the cases over the period of time.

TECHNOLOGY USED: MapReduce (in Java)

EXPECTED OUTPUT FORMAT:

<year> <certified> <certified-withdrawn> <withdrawn> <denied>
<certified %> <certified-withdrawn %> <withdrawn %> <denied %>

Query 6


Application count for each case status
600000
500000
400000
300000
200000
100000
0
Application count
2012
2013
2014
2015
2016
Year
Certified
Certified Withdrawn
Withdrawn
Denied

Query 7

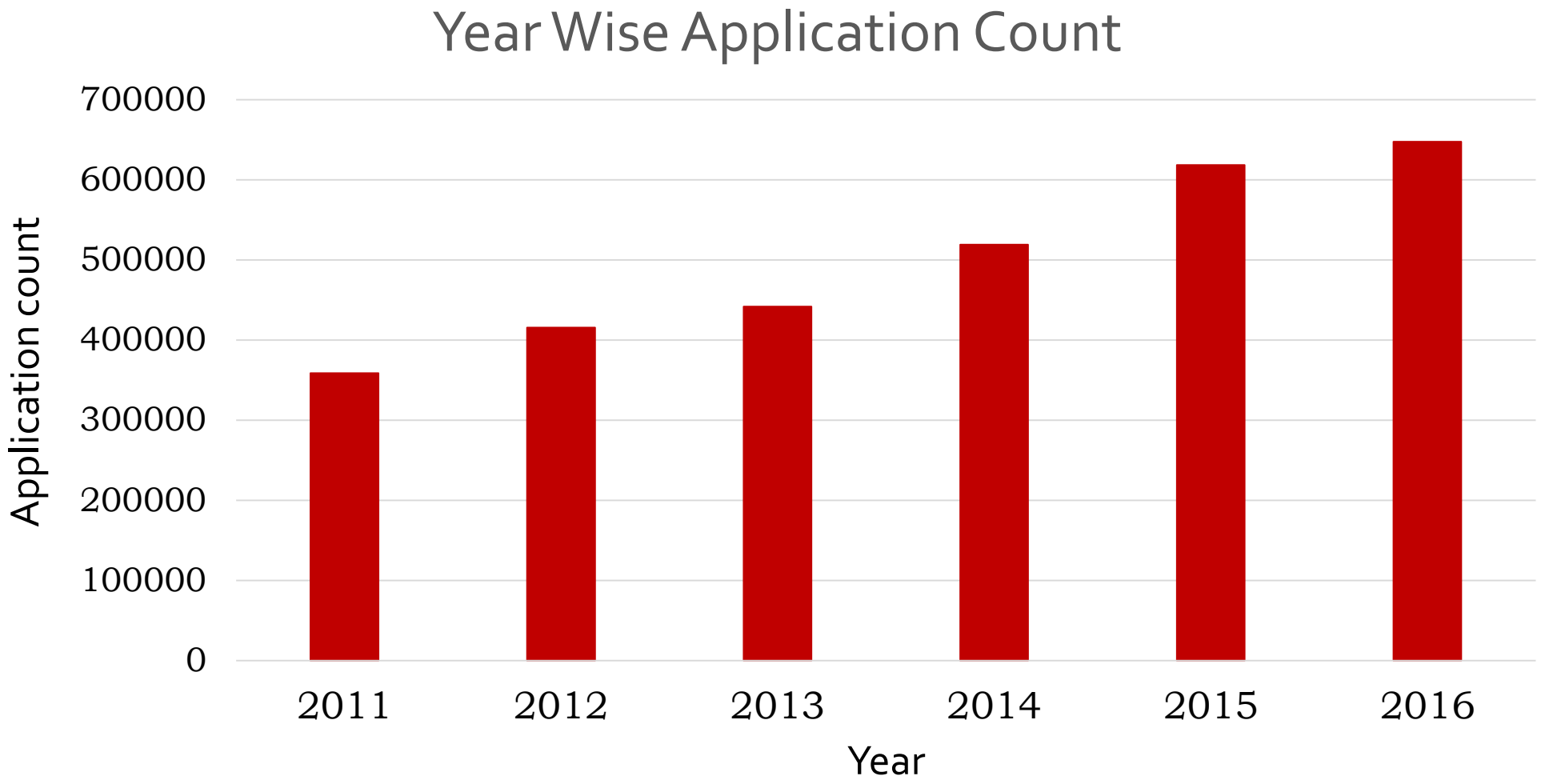
OBJECTIVE: To find the number of applications for each year. Use a bar graph to depict the same

TECHNOLOGY USED: Hive and LibreOffice Calc

EXPECTED OUTPUT FORMAT:

<soc name> <application count>

Query 7


Year Wise Application Count
700000
600000
500000
400000
300000
200000
100000
0
Application count
2011
2012
2013
2014
2015
2016
Year

Query 8

OBJECTIVE: To find the average prevailing wage for each job for each year (take part time and full time separate). Also, Arrange the output in descending order.

TECHNOLOGY USED: MapReduce (in Java)

EXPECTED OUTPUT FORMAT:

<job title> <year> <full time position> <average prevailing wage>

Query 9

OBJECTIVE: To find all the employers, along with the number of petitions, who have a success rate of more than 70% in petitions (total petitions filed 1000 OR more than 1000)

TECHNOLOGY USED: Pig

EXPECTED OUTPUT FORMAT:

<employer name> <certified> <certified-withdrawn> <total> <success rate>

Query 10

OBJECTIVE: To find all the job positions, along with the number of petitions, which have a success rate of more than 70% in petitions (total petitions filed 1000 OR more than 1000)

TECHNOLOGY USED: Pig

EXPECTED OUTPUT FORMAT:

<job title> <certified> <certified-withdrawn> <total> <success rate>

Query 11

OBJECTIVE: To export the result of Query 10 to MySQL database

TECHNOLOGY USED: Sqoop and MySQL

EXPECTED OUTPUT FORMAT:

<job title> <certified> <certified-withdrawn> <total> <success rate>

FIELD	TYPE	NULL	KEY	DEFAULT	EXTRA
job_title	varchar(60)	NO	PRI	NULL	
certified_count	int(11)	NO		NULL	
certified_withdrawn_count	int(11)	NO		NULL	
total_count	int(11)	NO		NULL	
success_rate	float	NO		NULL	

Conclusion

- Pig
  - ✓ Logic flow
  - ✓ Output of each step
- Hive
  - ✓ Data warehousing facilities
  - ✓ SQL-like interface
- MapReduce
  - ✓ Data - <key, value>
  - ✓ Complex aggregation operations

Thank you !