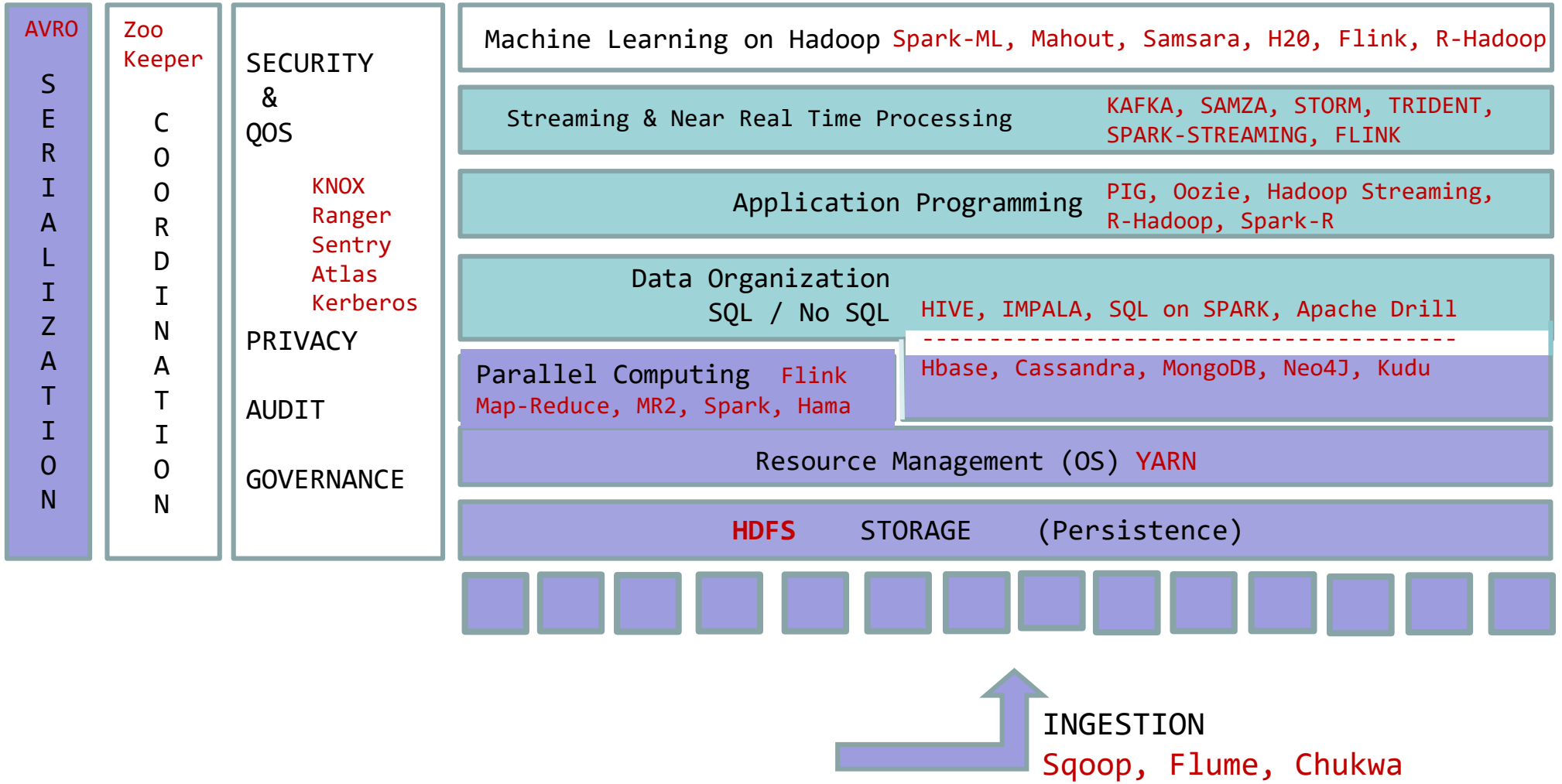
Inspire…Educate…Transform.

Business Intelligence on Hadoop

Our Focus: The open source big data ("Hadoop + Spark") ecosystem



- Hadoop has been around for about 10 years.
- Over the last 10 years, many innovations have been delivered, with technology breakthroughs that span a large space, from Apache Spark to Presto to Zeppelin.
- Hadoop got it start as a batch system, primarily used to store large volumes of data at a low cost.
- Now, thousands of enterprises have delivered Hadoop-powered applications for financial fraud detection, cybersecurity, inventory management, network troubleshooting, risk analysis, IoT… etc.
- Some even claim to use Hadoop to save lives: Cerner, a Cloudera customer, says it has developed an infection detection system that has saved more than 3,000 lives to date!

Source: https://www.forbes.com/sites/ciocentral/2017/08/15/hadoop-haters-gonna-hate/#19a6fa8a544e

Hadoop as an Analytics Platform

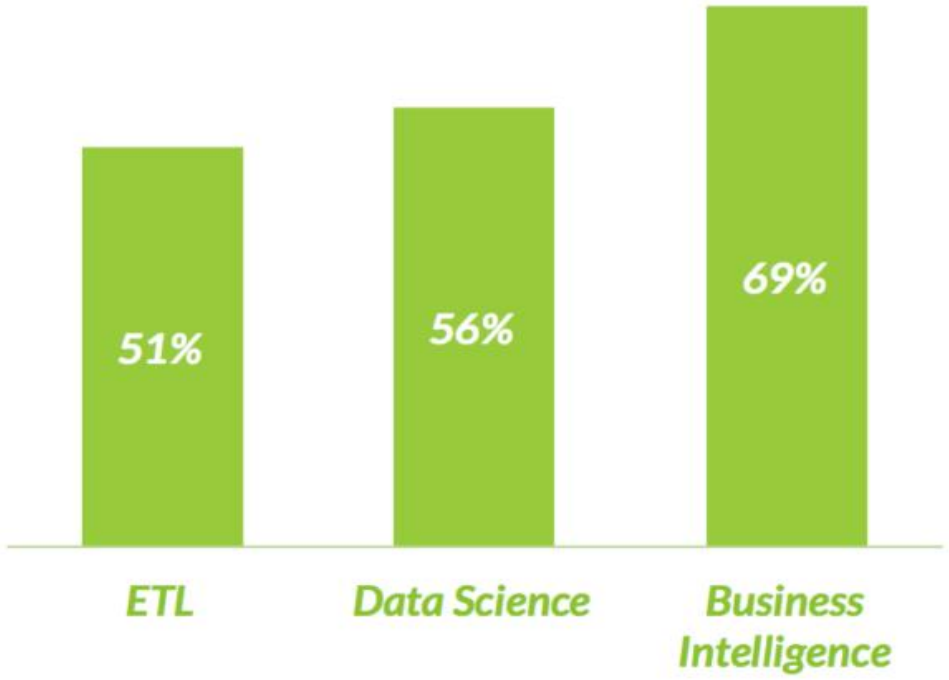
- Much of this change is due to the innovation, adoption, and commercialization of Hadoop.

- The limitations of traditional data infrastructures have led many organizations to move to Hadoop for not only new data use cases, but for day-today operational workloads as well. As Hadoop matures, enterprises are starting to use this powerful platform to serve more diverse workloads.

- Hadoop is no longer just a batch-processing platform for data science and machine learning use cases – it has evolved into a multi-purpose data platform for operational reporting, exploratory analysis, and real-time decision support.

- With the ongoing innovation of the SQL-on-Hadoop and in-memory data processing engines, Hadoop is now able to serve business-critical workloads in production. Hadoop is now ready to be the data source for business intelligence (BI) and online analytical processing (OLAP) workloads.

Hadoop Use Cases



Source: atscale.com/survey

Concepts

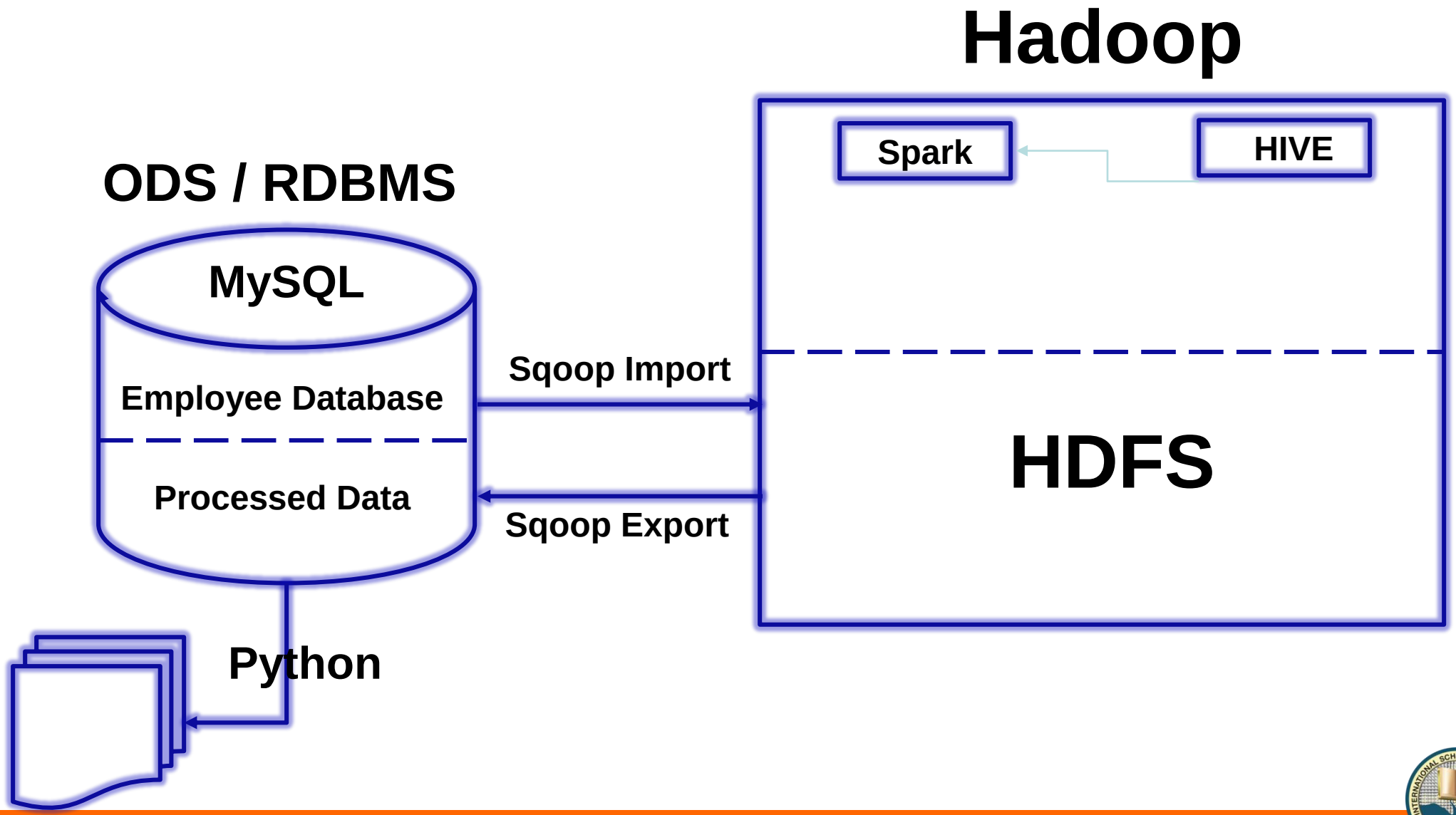
- Typical architecture a of BI System
- Operational Data Stores,
- Data Warehouse, Data marts, Business Intelligence,
- ETL (Extract, Transform, Load) Process,
- On-Line Transactional Processing (OLTP) vs On-Line Analytical Processing (OLAP).
- Different Ingestion tools/components
- Transformations (ETL vs ELT)
- MySQL, Sqoop, Hive, Spark, Python…

Problem Statement

- Given a large employee database with data scattered in 6 different tables
- Create a single large table/file with the corresponding details of all active employees and two aggregated tables one grouped based on department and other grouped based on department and gender.
- Derive descriptive measures on

1) Age vs Tenure
2) Age vs Salary
3) Tenure vs Salary
4) Employee's age distribution
5) Employee's tenure distribution
6) Employee's salary distribution
7) Male, Female counts in each department
8) Male and Female salary distribution

Process flow



Database Schema

employees
emp_no
birth_date
first_name
last_name
gender
hire_date
last_modified

dept_emp
seq_no
emp_no
dept_no
from_date
to_date
last_modified

titles
seq_no
emp_no
title
from_date
to_date
last_modified

departments
dept_no
dept_name
last_modified

dept_manager
seq_no
dept_no
emp_no
from_date
to_date
last_modified

salaries
seq_no
emp_no
salary
from_date
to_date
last_modified

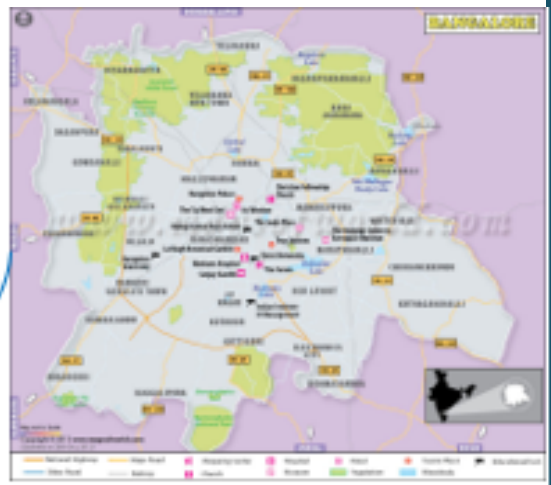
Detailed Steps

- Employee database with more than 300,000+ employees on a MySQL server as ODS/OLTP/Source.

1. Import this database (all these tables) to HDFS using Sqoop.
2. Incremental Imports.
   a. Add some rows to one/more table(s).
   b. Do an incremental import to HDFS using Sqoop.
   c. Modify some rows in one/more table(s).
   d. Do an incremental import to HDFS using Sqoop.
3. Create Hive Tables (Define Schema) for the above data in HDFS.
   This enable us to access the above data from all those big data components which have access to hive meta store, one don't have to define the schema again while working with this data .
4. Process the above imported data using Spark.
   a. Access all the hive tables from Spark and create intermediate data frames.
   b. Process those data frames using Spark SQL or data frame operations to gather all the relative fields which are going to give us the insights we are looking for.
5. Aggregate data using Spark SQL.
   What ever data we get from step 4 we will try to aggregate it based upon either department or by both department and gender
6. Export the data created in 5 to MySQL using Sqoop Export.
7. Draw plots/visualizations using Python and with the above data available in MySQL.

INSOFE

Inspire...Educate...Transform.



HYDERABAD

Office and Classrooms

Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Office

Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Classroom

KnowledgeHut Solutions Pvt. Ltd., Reliable Plaza,
Jakkasandra Main Road, Teacher's Colony, 14th Main
Road, Sector – 5, HSR Layout, Bengaluru - 560102

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.