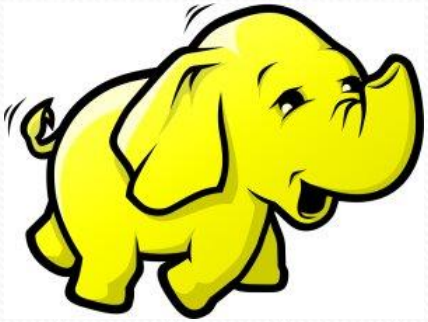
Hadoop Application in Health Care

Luyao Zhang

INLS770 Presentation

What is Hadoop

An open-source software framework for **storing data** and **running applications** on **clusters** of commodity hardware



Big Data - 3V

- Volume
- Variety
- Velocity

Hadoop - Core components

- HDFS（Hadoop Distributed File System）
- MapReduce Computing System

Hadoop – Cluster and Node

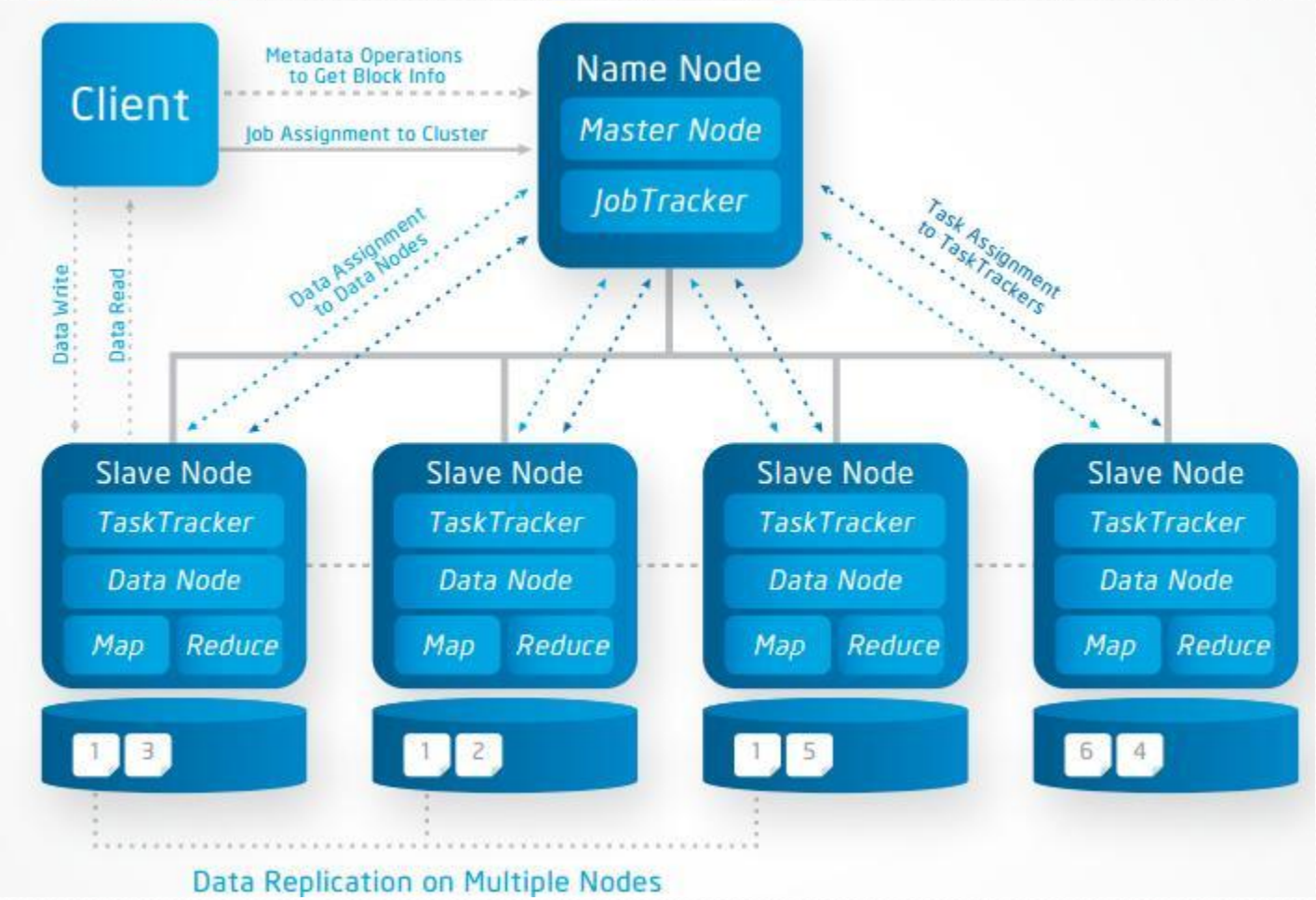
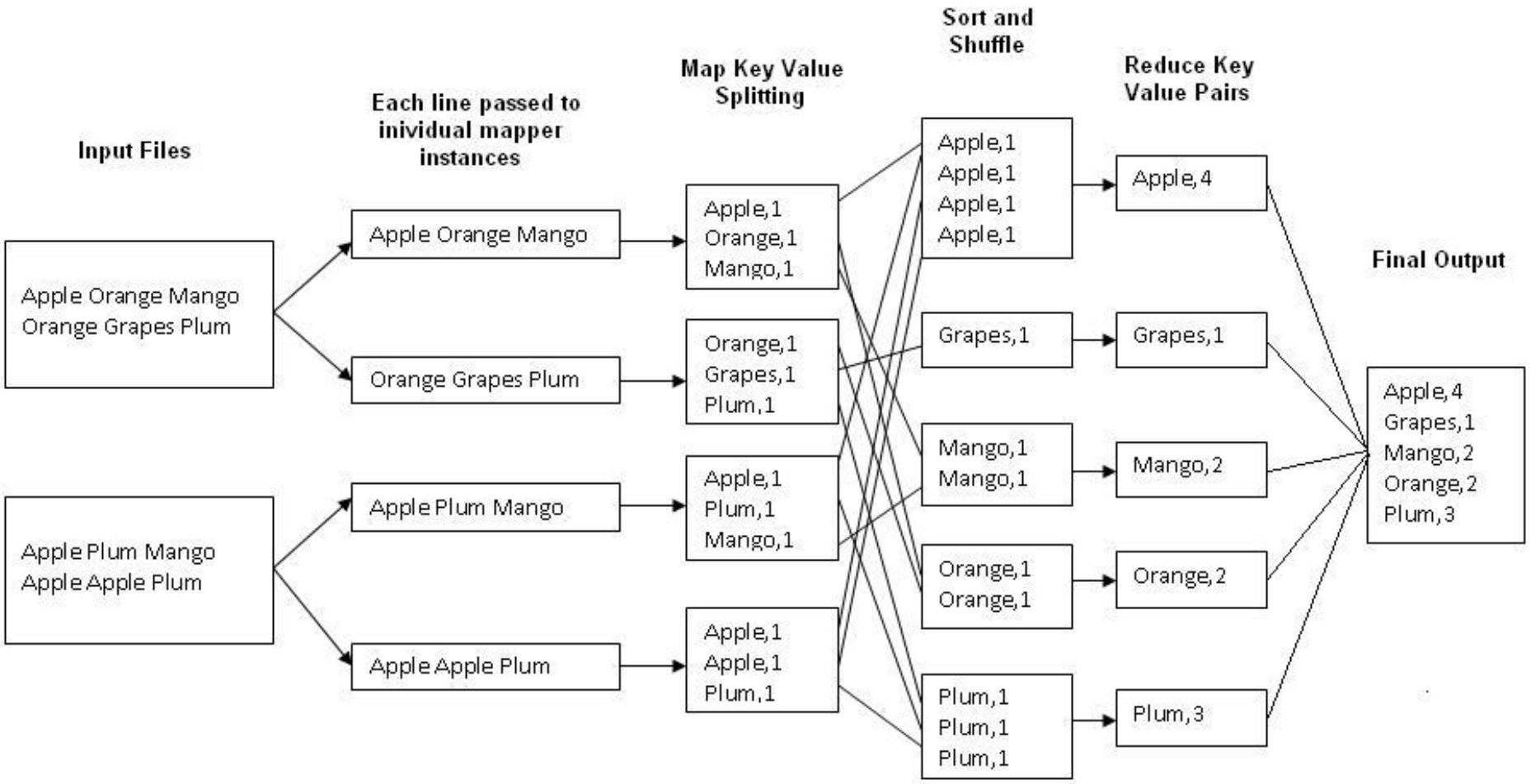


Image source: http://donatz.info/hadoop-cluster-architecture/hadoop-cluster-architecture-creative-on-inside-with-and-6/

Hadoop – MapReduce



Hadoop - Ecosystem

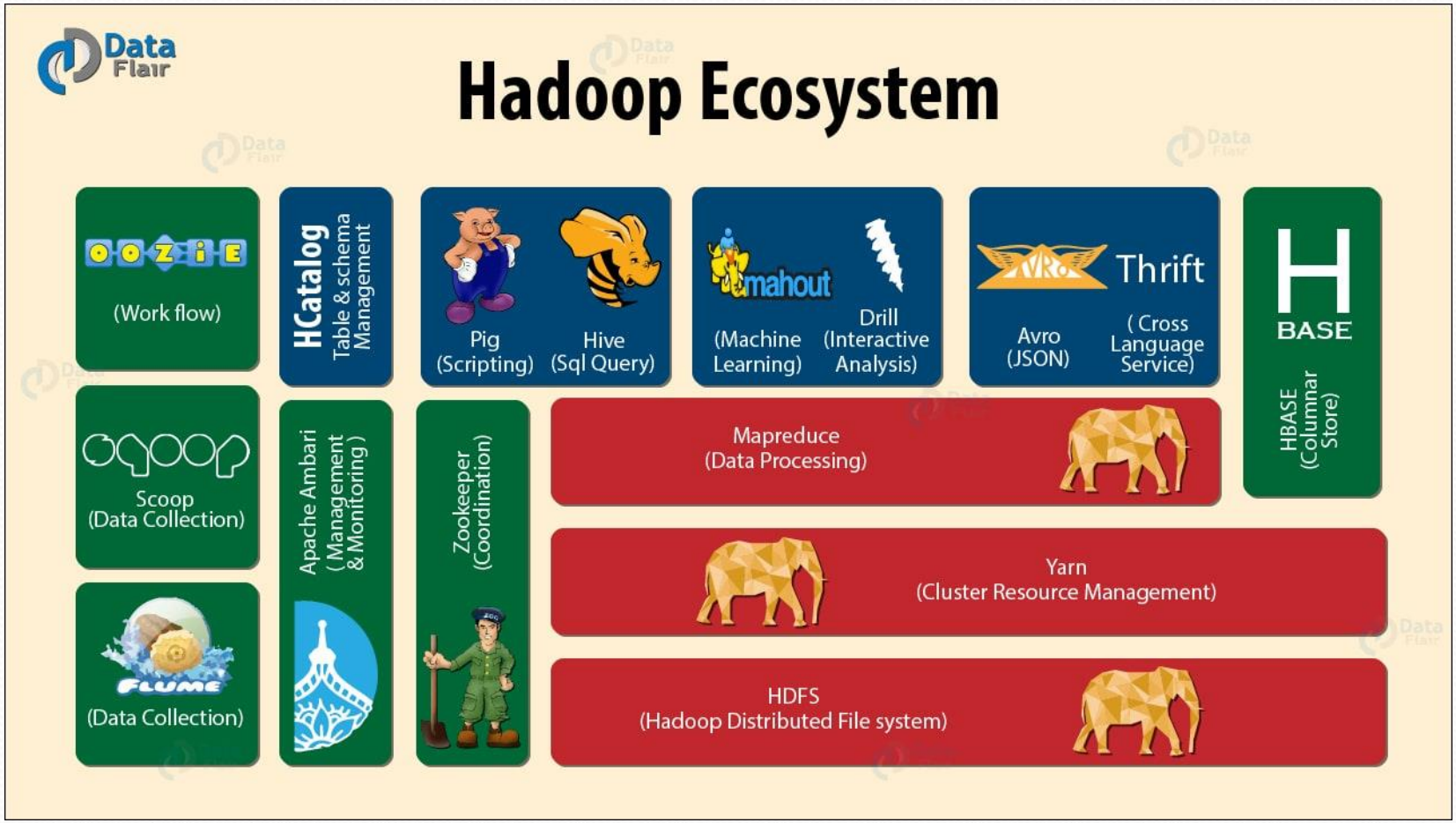
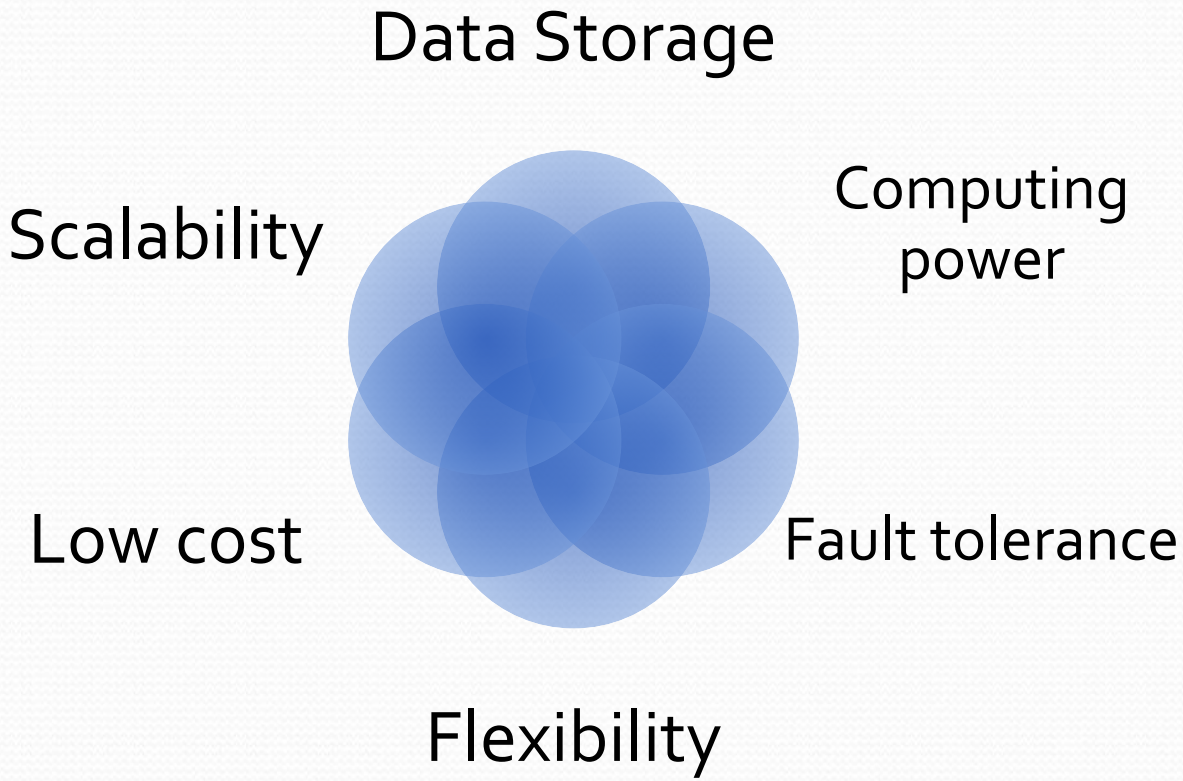
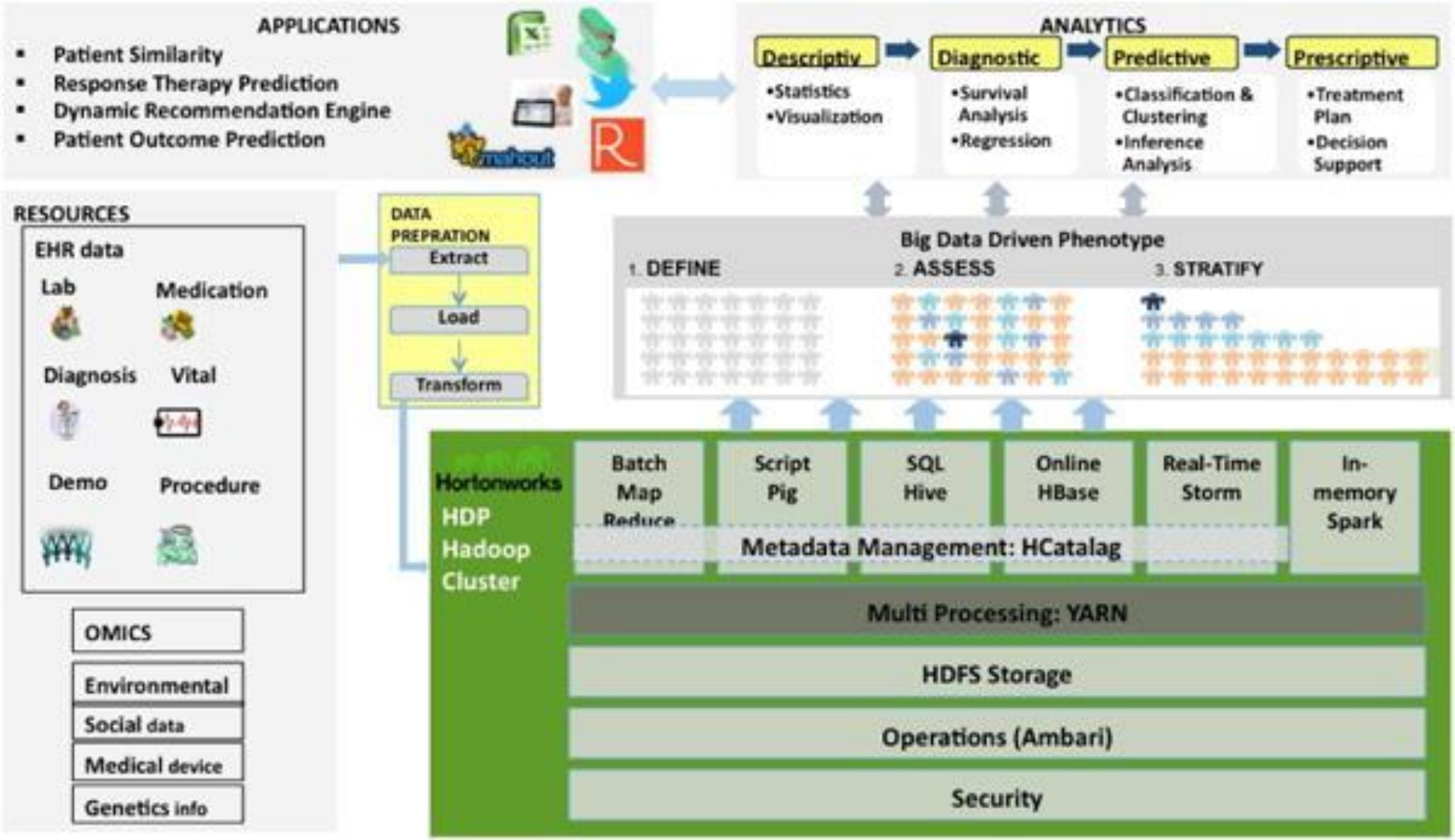


Image source: https://data-flair.training/blogs/hadoop-ecosystem-components/

Why Hadoop


Data Storage
Computing power
Fault tolerance
Flexibility
Low cost
Scalability

Hadoop Data Platform in Mayo Clinic



Cohort creation

- Patients with heart failure diagnosis event with at least one EF (Ejection Fraction) value within three months of the heart failure diagnosis date
- Data sources: four large datasets from different clinical systems
- Tool: **Apache Pig**

Apache Pig

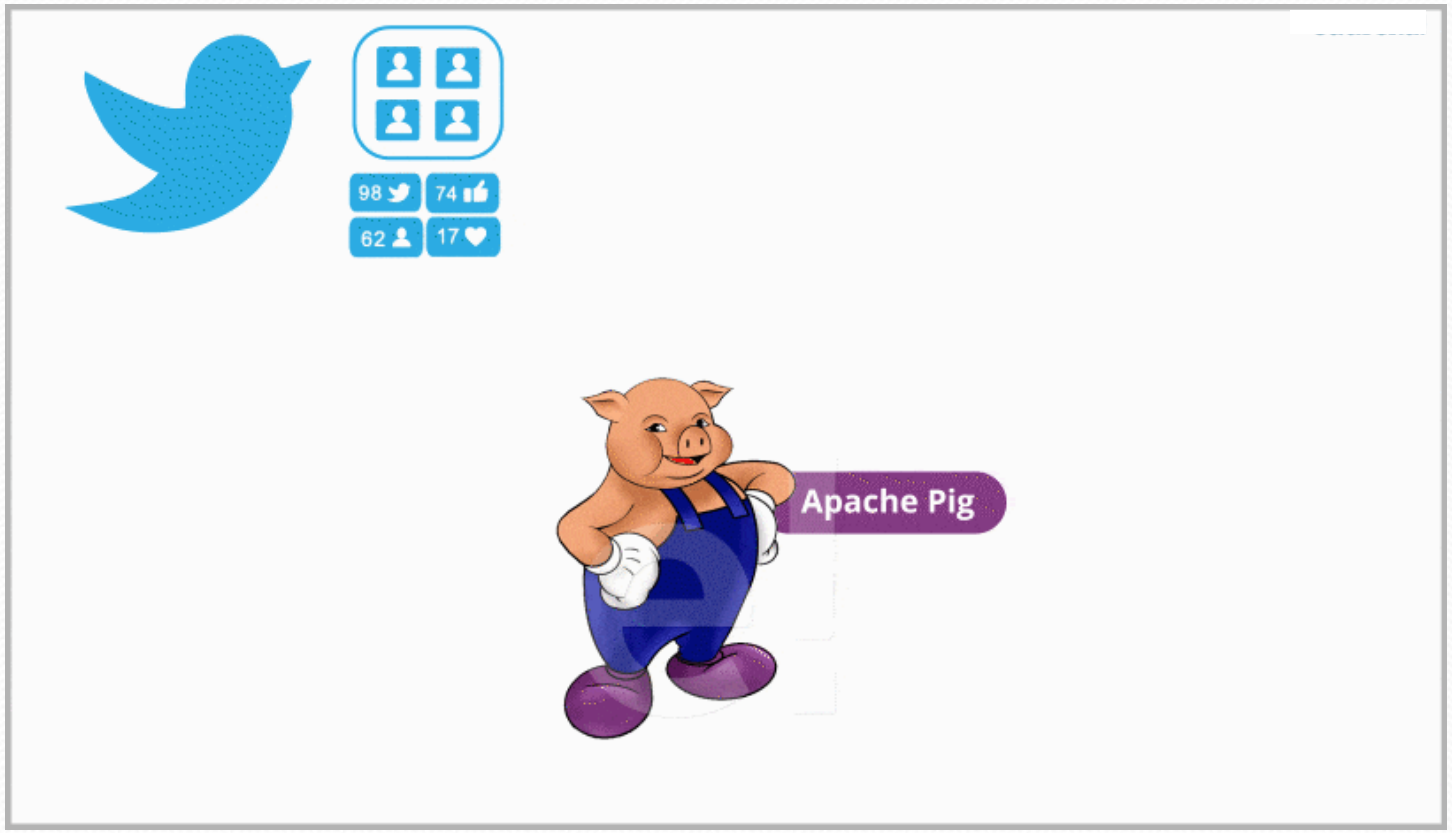


Image source: Edureka

Challenges

- Implementations of analytical methods in healthcare
- Pipeline automation
- Restricted to batch processing, not for real-time processing

References

1. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: a literature review. Biomedical informatics insights, 8, BII-S31559.
2. Rallapalli, S., Gondkar, R. R., & Ketavarapu, U. P. K. (2016). Impact of processing and analyzing healthcare big data on cloud computing environment by implementing hadoop cluster. Procedia Computer Science, 85, 16-22.
3. Panahiazar, M., Taslimitehrani, V., Jadhav, A., & Pathak, J. (2014, October). Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 790-795). IEEE.
4. Peek, N., Holmes, J. H., & Sun, J. (2014). Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. Yearbook of medical informatics, 9(1), 42.
5. Agarwal, P., & Owzar, K. (2014). Next generation distributed computing for cancer research. Cancer informatics, 13, CIN-S16344.