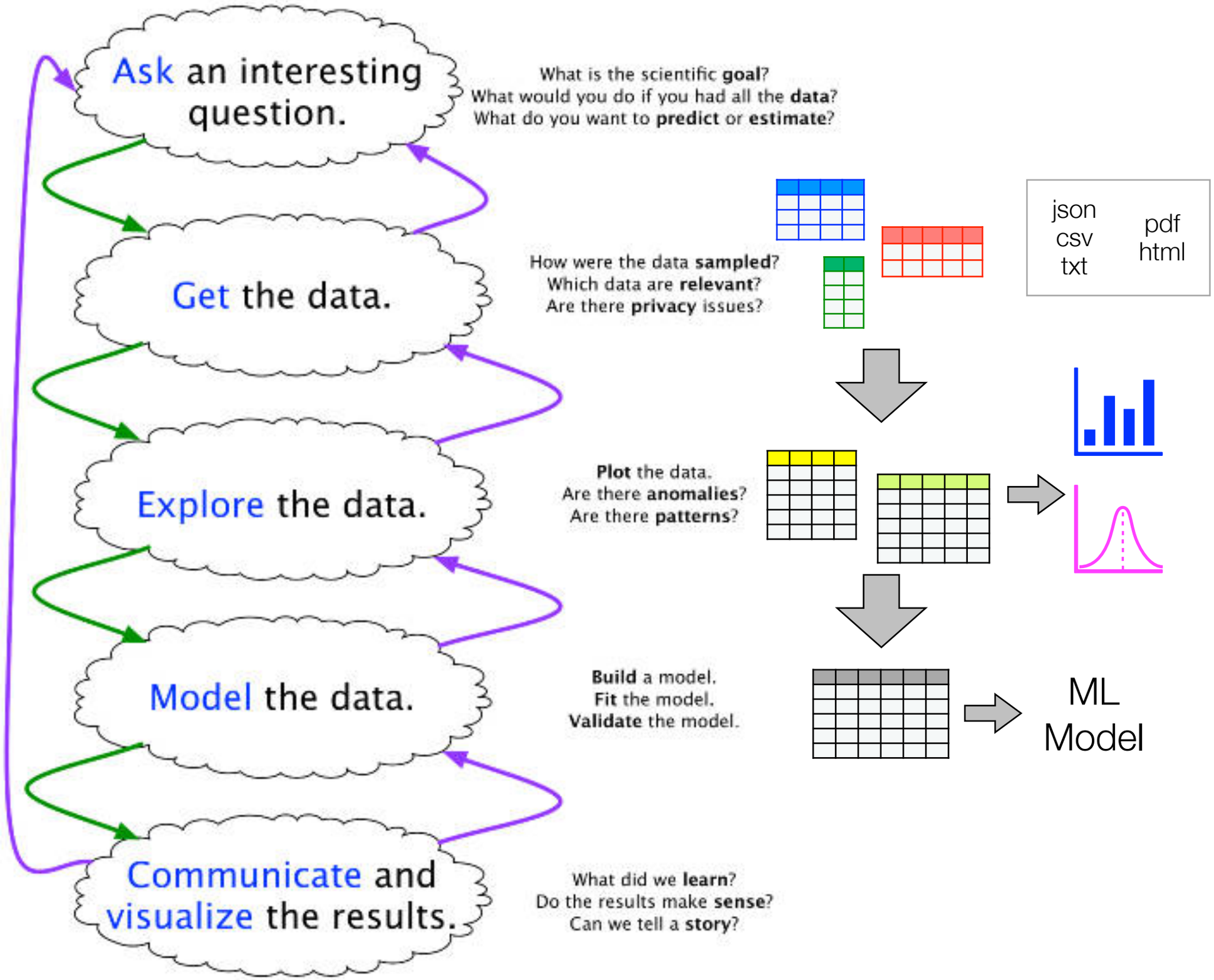
2110403 - Introduction to Data Science and Data Engineering

Introduction to Data Engineering

Asst.Prof. Natawut Nupairoj, Ph.D.
Department of Computer Engineering
Chulalongkorn University
natawut.n@chula.ac.th

CHULA ENGINEERING
Foundation toward Innovation

The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Source: https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html

What is Data Engineering? (Wikipedia)

- Data engineering refers to the building of systems to enable the collection and usage of data
- This data is usually used to enable subsequent analysis and data science, which often involves machine learning
- Making the data usable usually involves substantial compute and storage, as well as data processing

The Data Engineer Job Market in 2024 [Research on 1,000 Job Postings]

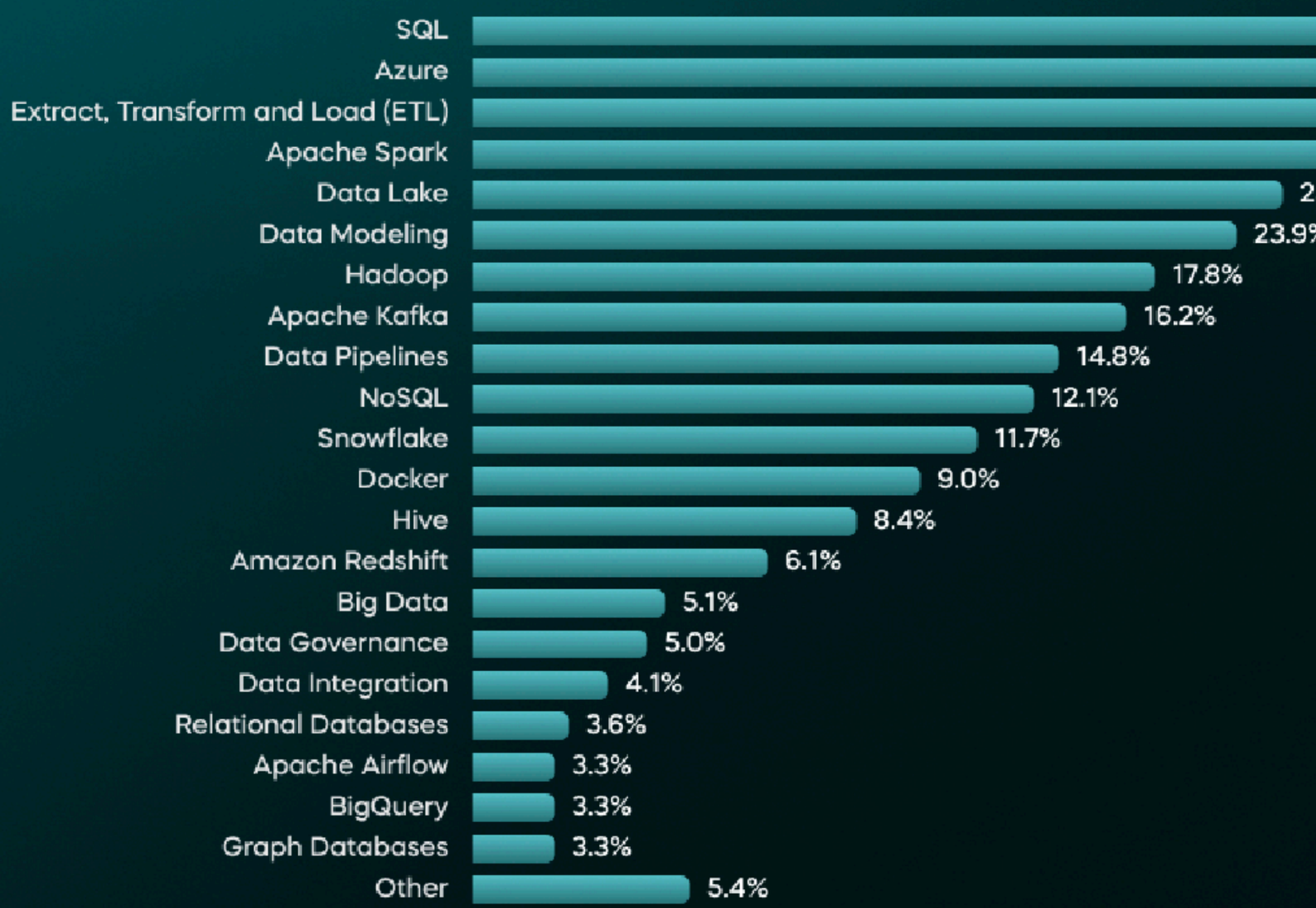
Sophie Magnet • 11 Sep 2024 • 18 min read

With the emergence of ChatGPT, recent years have seen concerns about how artificial intelligence (AI) could lead to job losses in data-related fields like data science and data engineering. There was speculation that AI could replace these roles because of its capacity to automate tasks traditionally performed by humans.

Source: https://365datascience.com/career-advice/data-engineer-job-market/

365

Data Engineering Skills



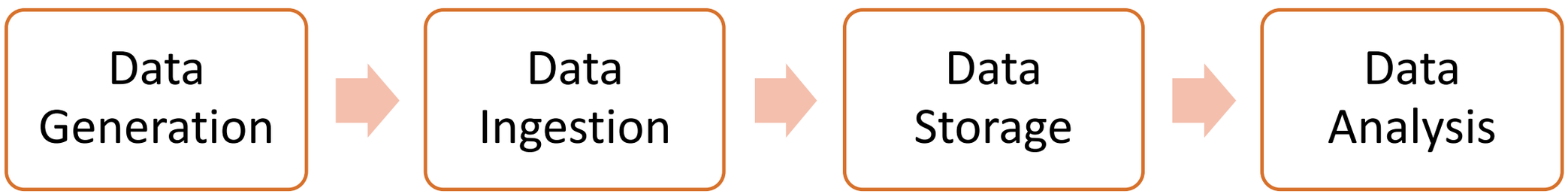
Data Engineering Skill Groups

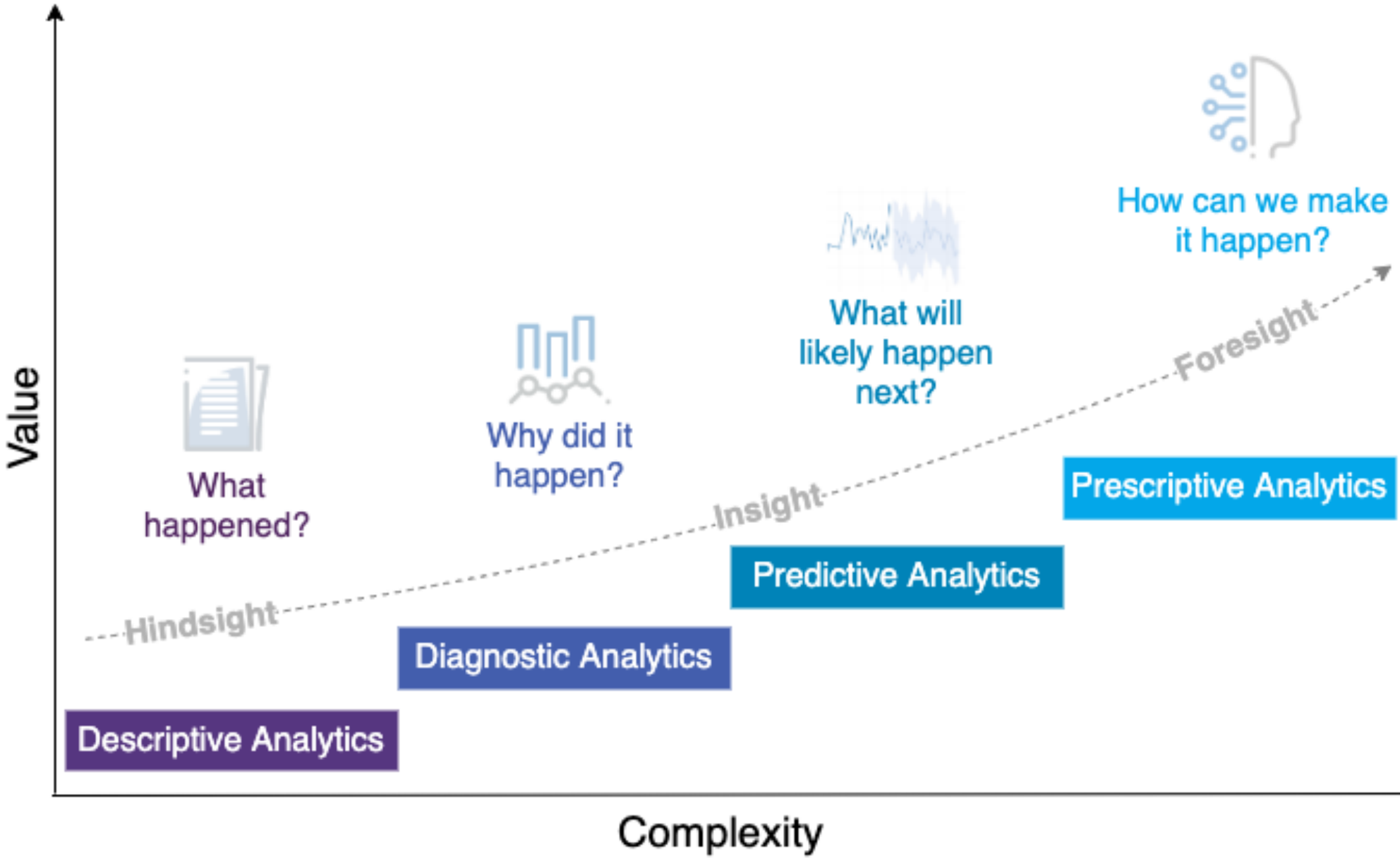
- Database Management
- Big Data Technologies
- Data Warehouse / ETL
- Cloud
- Data Pipeline and Workflow Management
- Data Governance
- Containerization (e.g. docker, k8s)

CHULA ENGINEERING
Foundation toward Innovation

Data Pipeline Analogy

Data Lifecycle





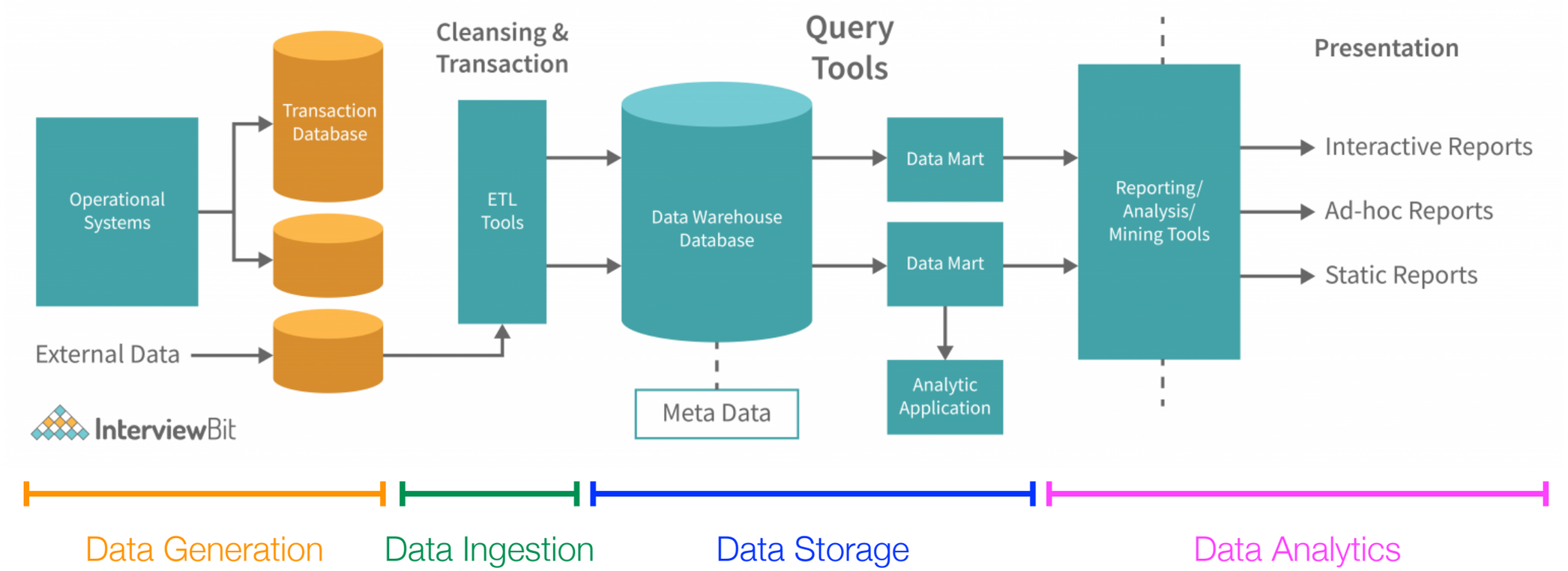
The 4 types of Data Analytics (Adapted from Davenport & Harris 2007 / Gartner 2012)

Source: https://www.scalesology.com/post/data-analytics-can-turn-hindsight-into-foresight

Data Analytics Simplified

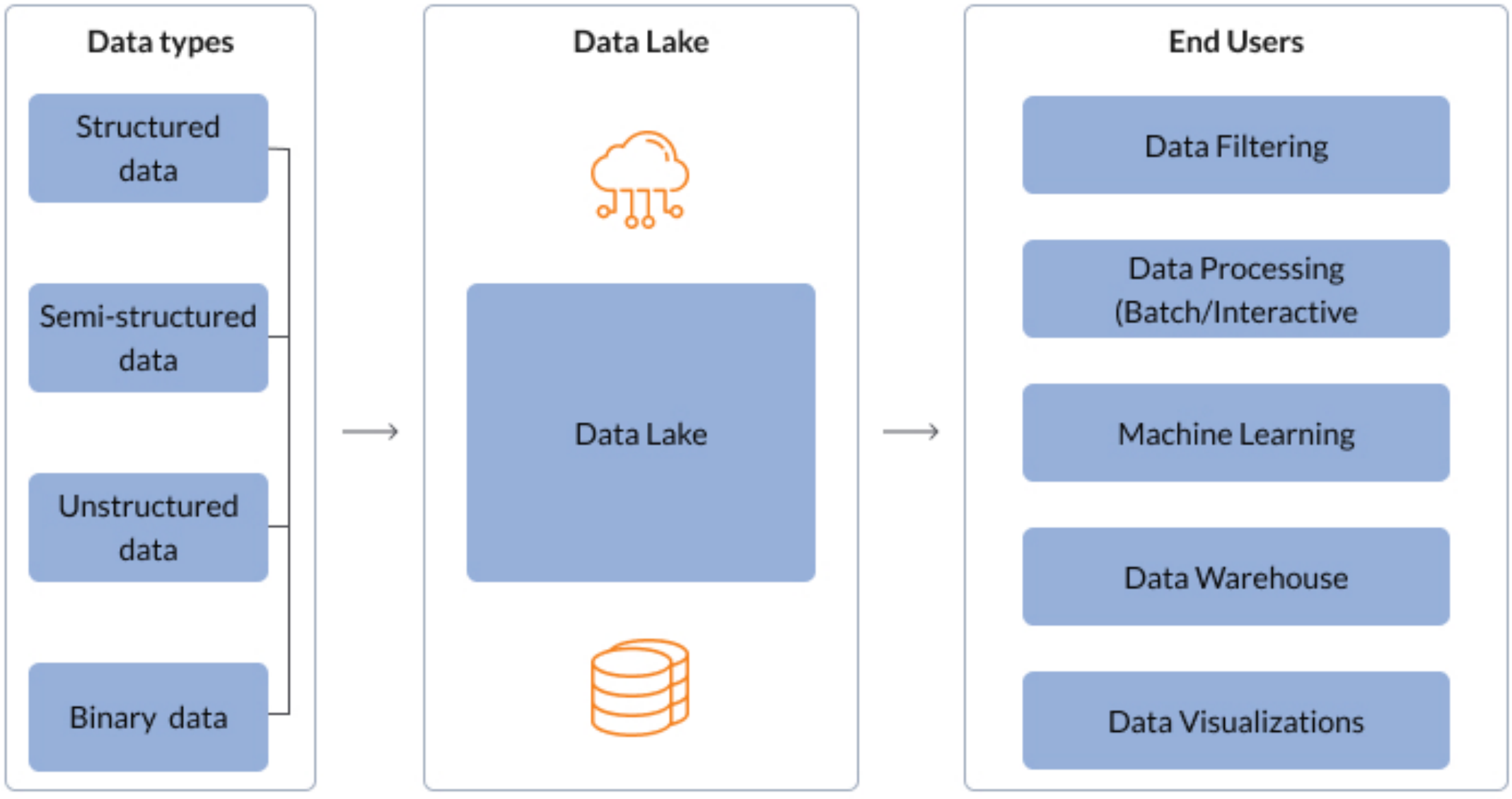
Descriptive	“A.Natawut drinks about 1 cup of coffee a day”
Diagnostic	“Number of cups that A.Natawut drinks depends on number of meetings he has each day”
Predictive	“Tomorrow, A.Natawut has 2 meetings. It is very likely that A.Natawut will drinks 2 cups.”
Prescriptive	“Inform secretary to prepare one cup in the morning and one in the afternoon for A.Natawut”

Data Warehouse + BI = Diagnostic Analytics Tool



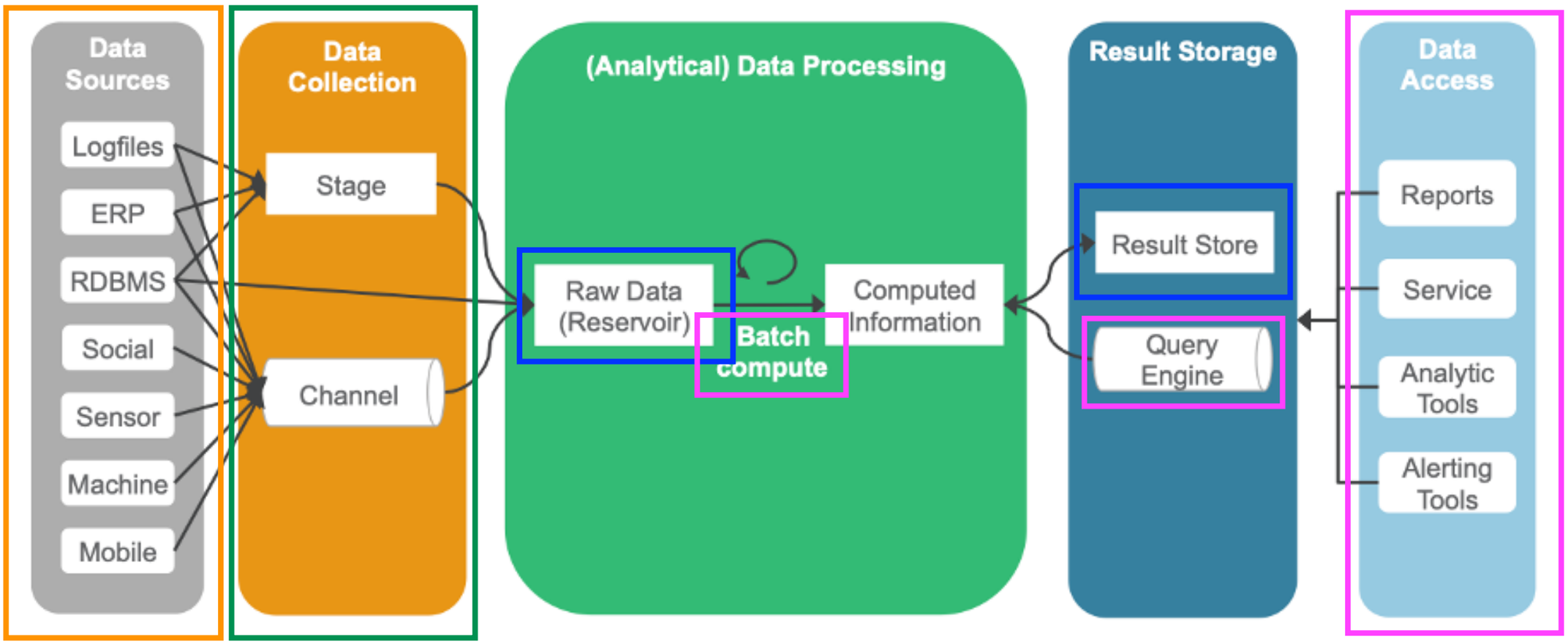
Source: “Data Warehouse Architecture – Detailed Explanation”, https://www.interviewbit.com/blog/data-warehouse-architecture/

Data Lake Architecture



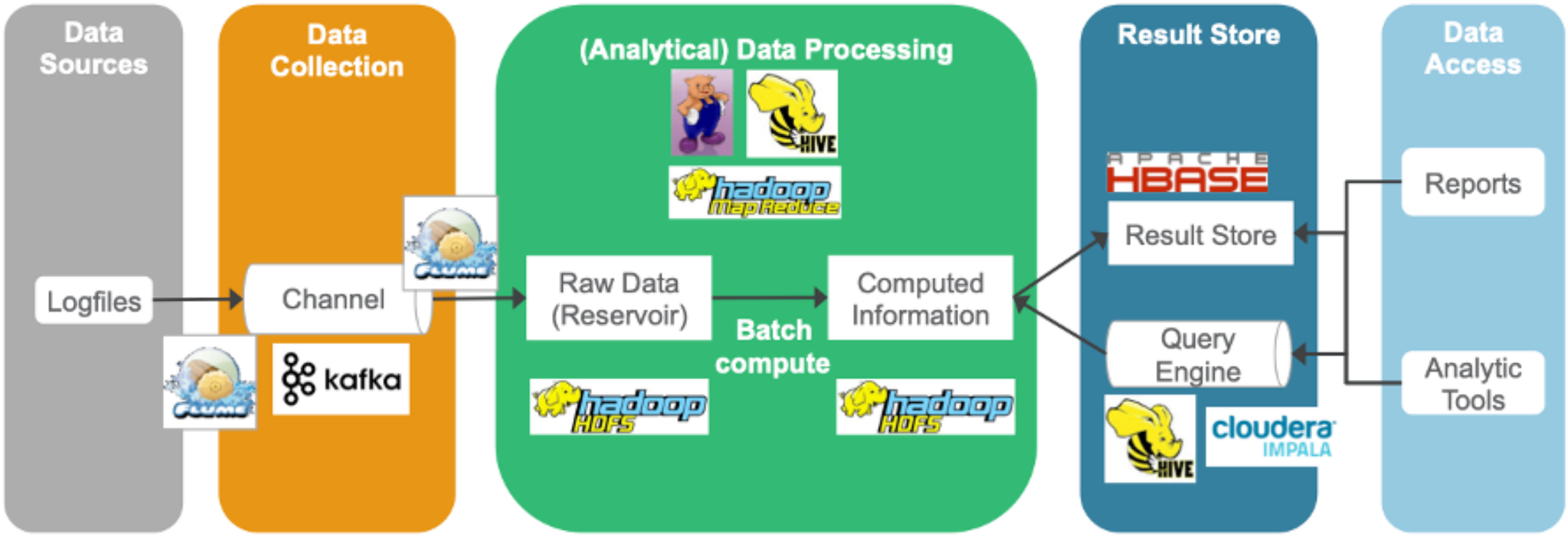
Source: https://www.n-ix.com/data-lake-vs-data-warehouse/

Simple Big Data Analytic Architecture

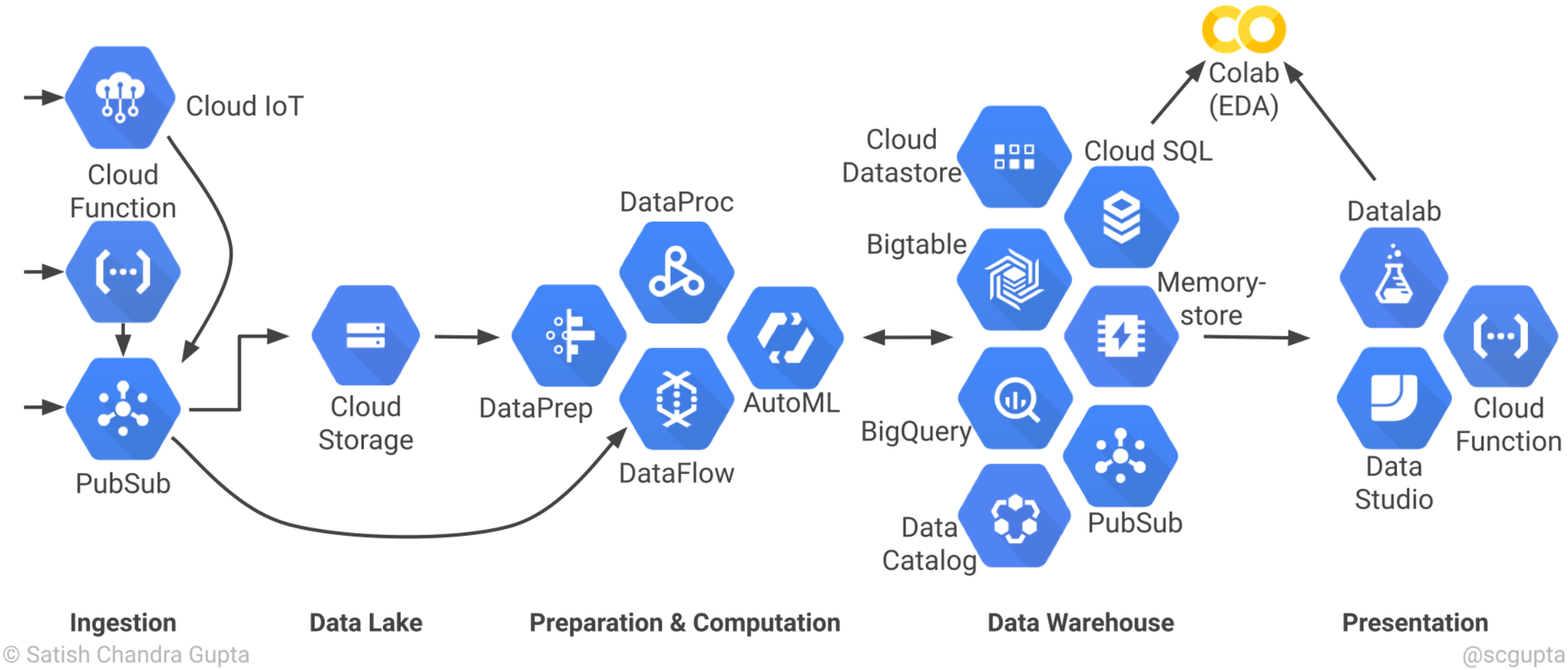


Data Generation Data Ingestion Data Storage Data Analytics

Example: Facebook Data Pipeline (early days)



Cloud Serverless Architecture



Source: https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5

Conclusion

- Data engineering is a process that makes data usable e.g. ready for analytics or ML modeling

- Typical data lifecycle = Generation / Ingestion / Storage / Analysis

- Big Data is just a foundation for data processing; data analytics is also important as it creates business values

- Data warehouse and data lake are two major data processing architectures

References

- D. Reinsel, J. Gantz, and J. Rydning, “Data Age 2025: The Digitization of the World From Edge to Core,” International Data Corporation, 2018.

- A. Menon, “Big data@ facebook,” in Proceedings of the 2012 workshop on Management of big data systems, 2012, pp. 31–32.

- J. Warren and N. Marz, Big Data: Principles and best practices of scalable realtime data systems. Simon and Schuster, 2015.

- “Data Warehouse Architecture – Detailed Explanation”, https://www.interviewbit.com/blog/data-warehouse-architecture/

- H. Leano, “How to Evaluate Data Pipelines for Cost to Performance”, https://www.databricks.com/blog/2020/11/13/how-to-evaluate-data-pipelines-for-cost-to-performance.html