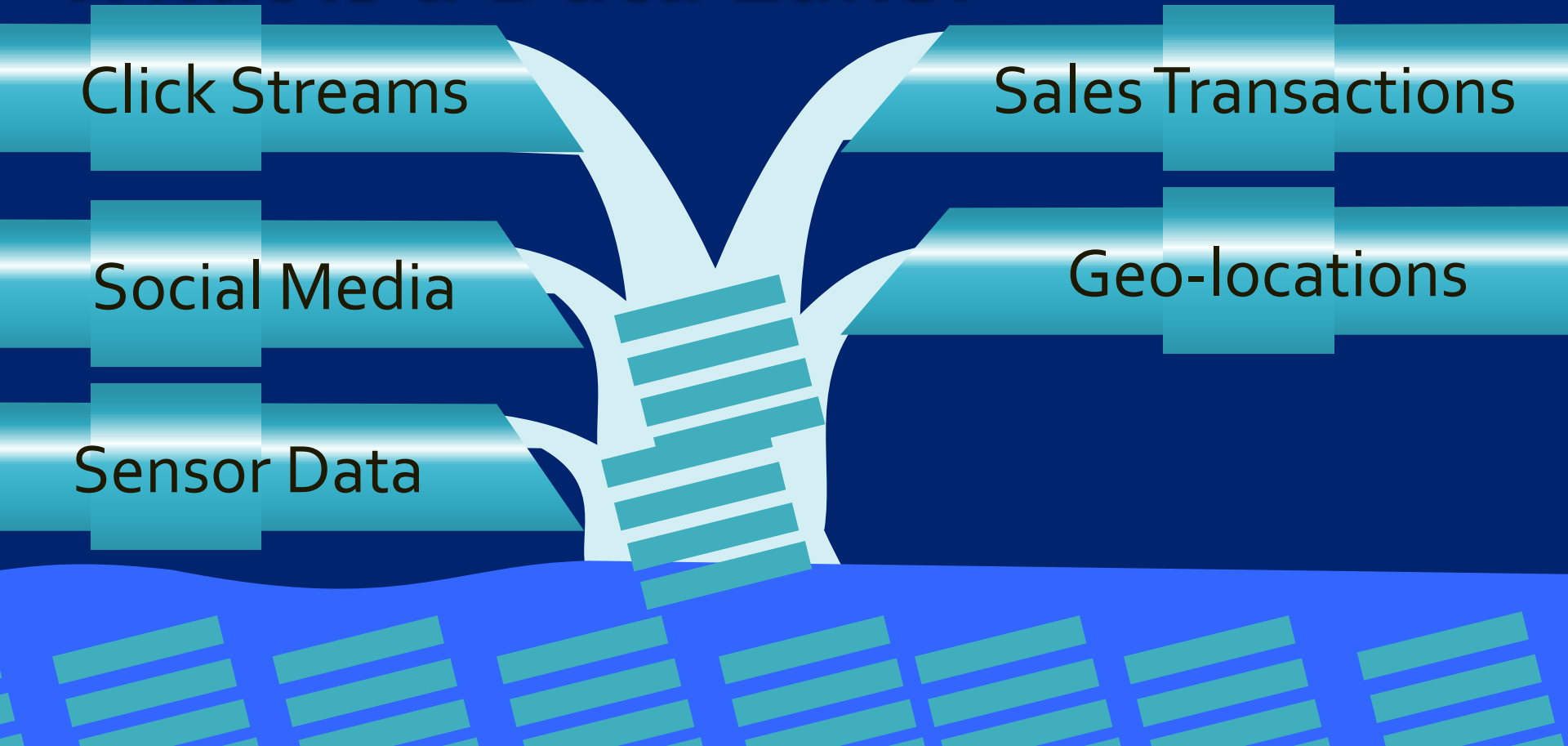
Understanding Data Lakes

After this video you will be able to…

- Describe how data lakes enable batch processing of streaming data
- Explain the difference between “schema-on-write” and “schema-on-read”
- Organize data streams, data lakes and data warehouses on a spectrum of big data management and storage

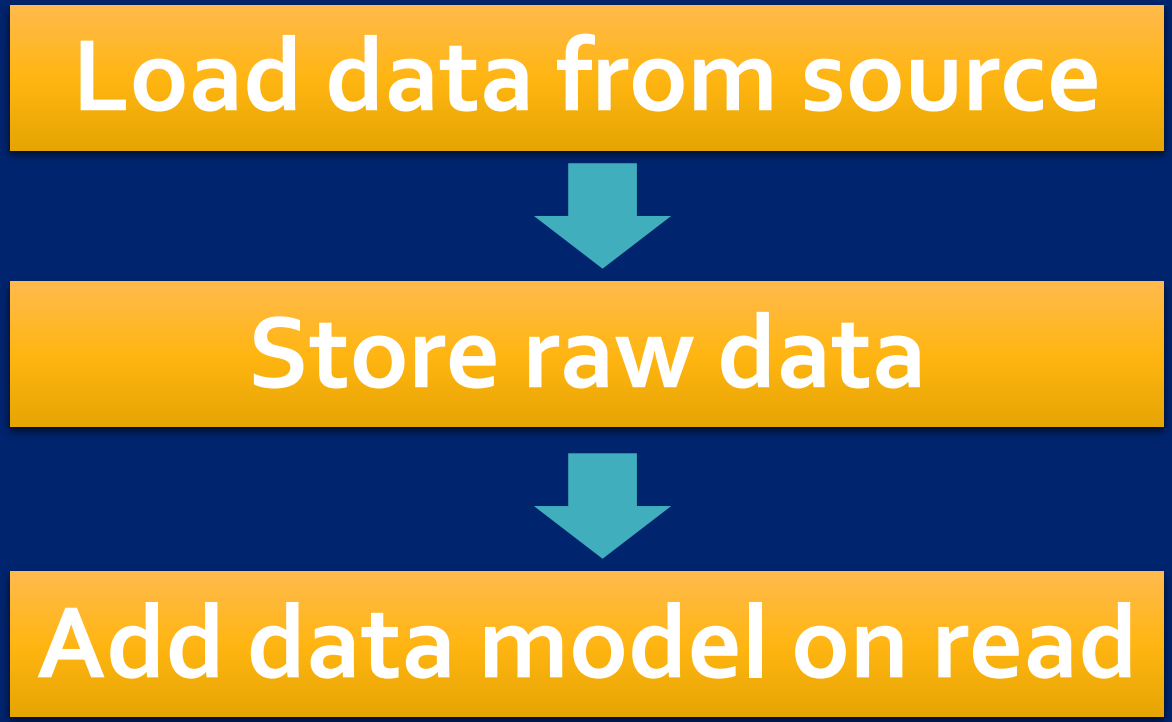
What is a Data Lake?



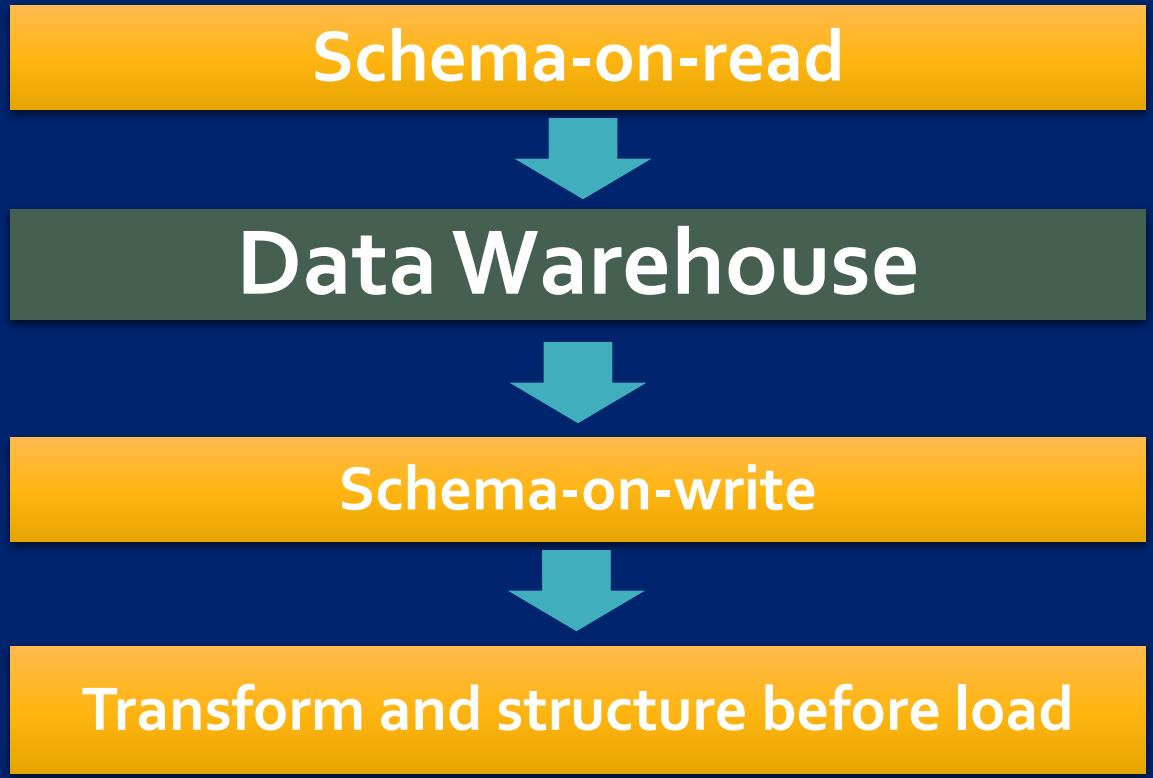
"If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

- James Dixon, CTO
Pentaho Corporation

How a Data Lake Works

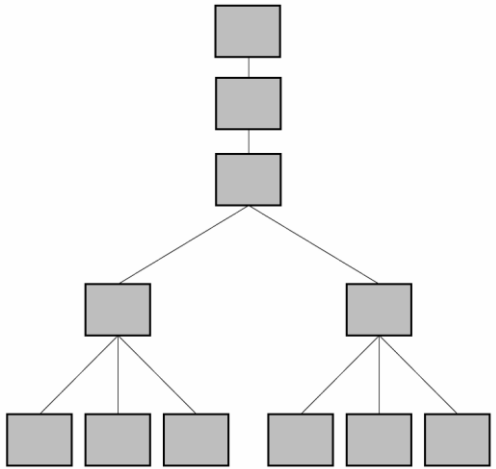


Schema-on-Read Approach

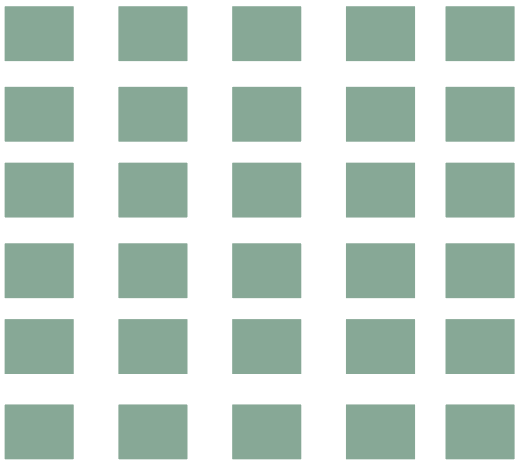


Data Lake vs. Data Warehouse

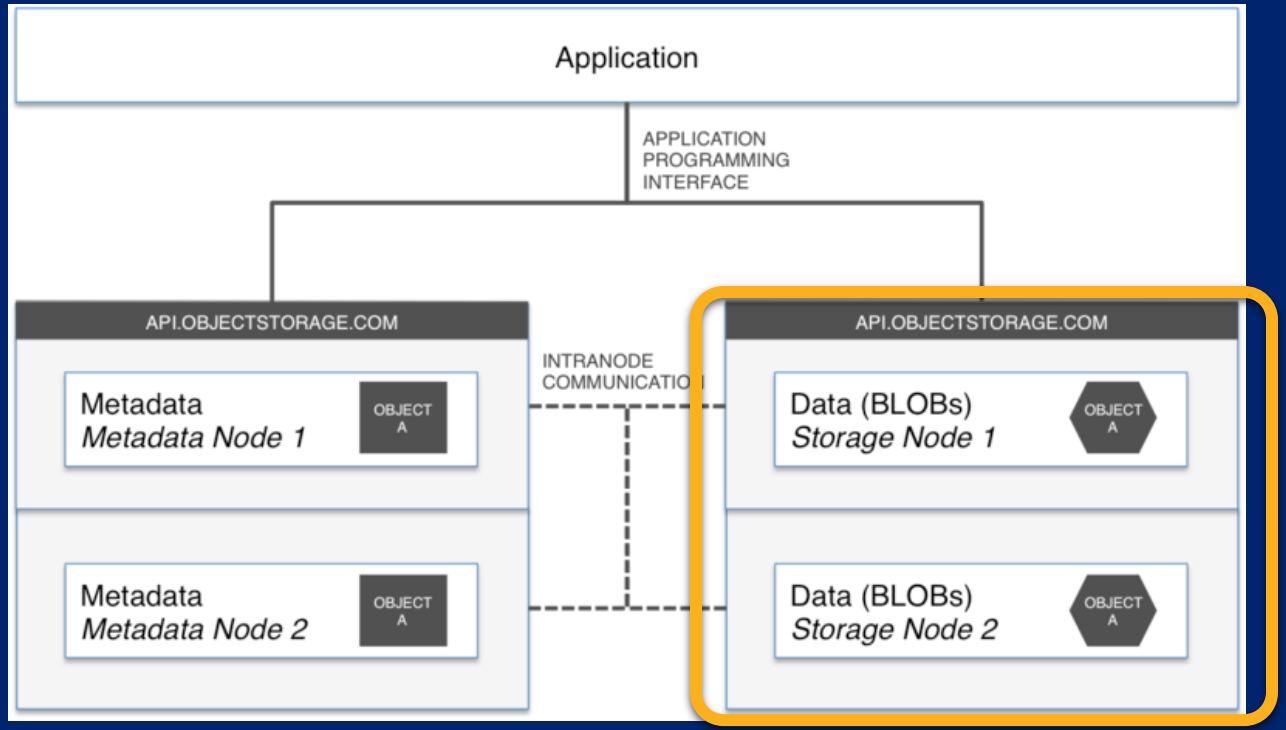
Data Warehouse: Hierarchical File System



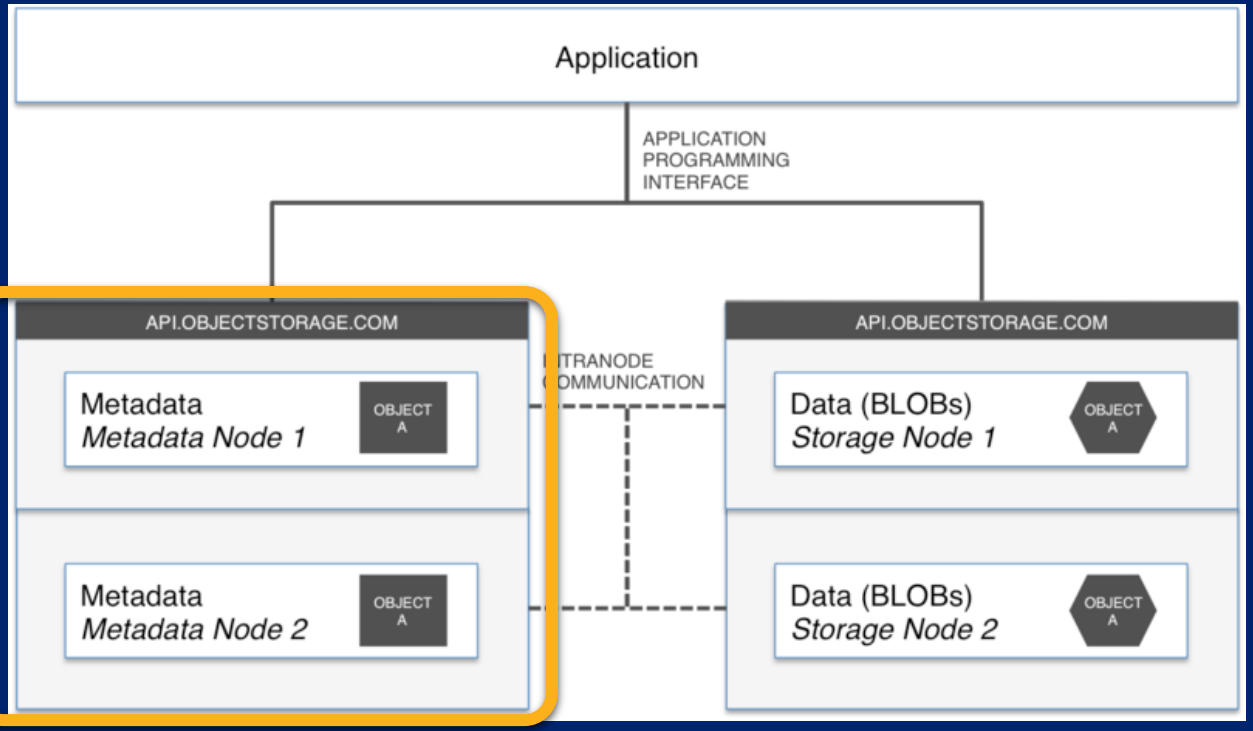
Data Lake: Object storage



Data Lake Object Storage



Data Lake Object Storage



Data Lakes Summary

- A Big Data Storage Architecture
- Collects all data for current and future analysis
- Transforms data format only when needed
- Supports all types of big data users
- Infrastructure components which evolve over time based on application-specific needs