Data Warehouse or Data Lake: Which one is right for you?

Traditional Data Warehouse		Data Lake
Structured	**Types of data**	Structured, semi-structured, unstructured, and raw
Expensive for large volumes of data	**Cost effectiveness**	Designed for low cost (typically on commodity hardware)
Business users	**Intended users**	Data scientists and advanced users
Business Intelligence, reporting, dashboards and visualizations	**Purpose**	Not always predefined, but may be leveraged for machine learning and AI
Expensive and time consuming to change due to schema design implications	**Flexibility**	More flexible and easier to update due to lack of rigid schema definition, but may introduce challlenges due to number of small files or complexity of Hadoop File System (HDFS)
Data is processed and organized into a single schema before storing in the data warehouse	**Ingestion process**	Raw and unstructured data can be directly ingested into the data lake
Data is already structured and can be consumed by business intelligence tools	**Analysis process**	Data is organized on demand at the time of analysis. Schema on read rather than write. Datalake is horizontally scalable.
Centralized	**Data storage**	Centralized
Relational (e.g. SQL)	**Schema type**	Undefined schema
Data is collected from multiple relational sources	**Data sources**	Collect data from any source and in any format and provide near real-time ingestion and streaming pipelines
Medium	**Data volume**	High
Established definitions of data structures and schemas	**Data definition**	Requires data to be cleaned, enriched and joined with metadata for consumption and or feeding into Data Warehouse

FACTION®