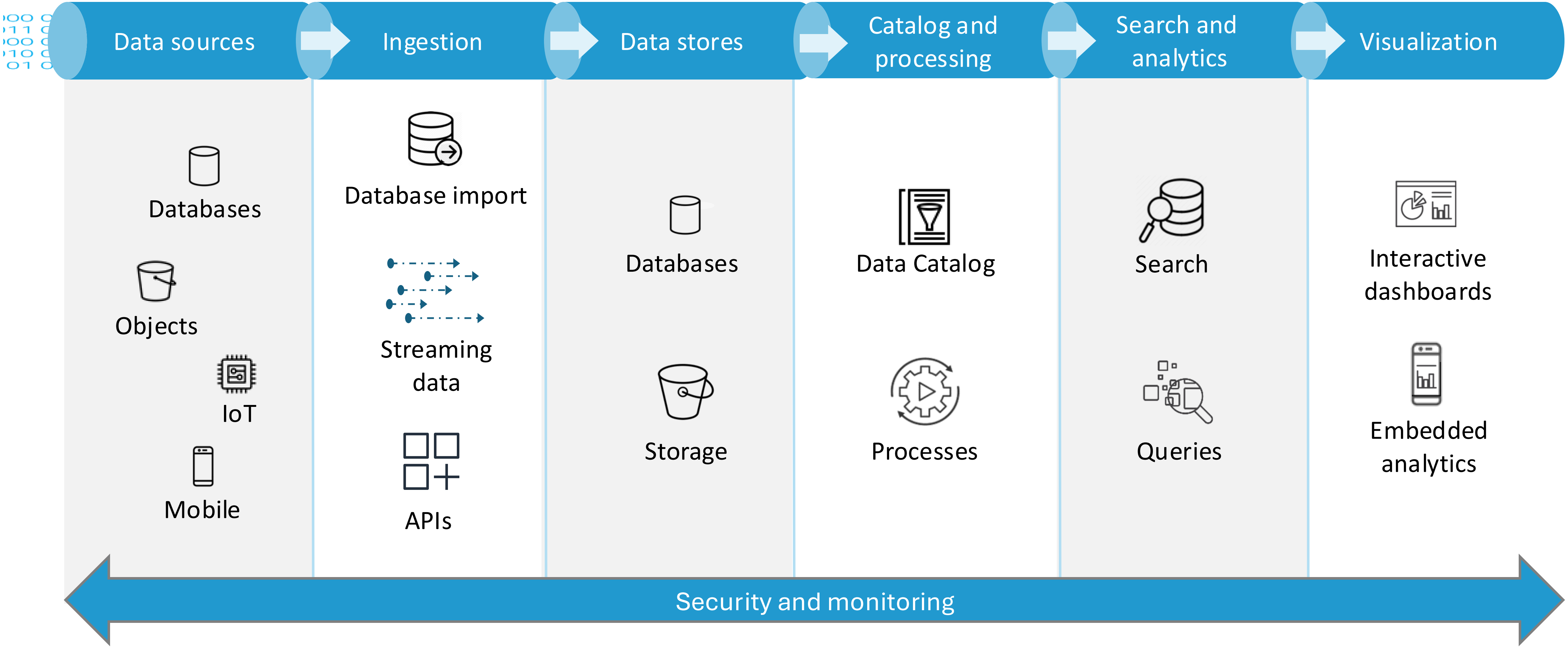
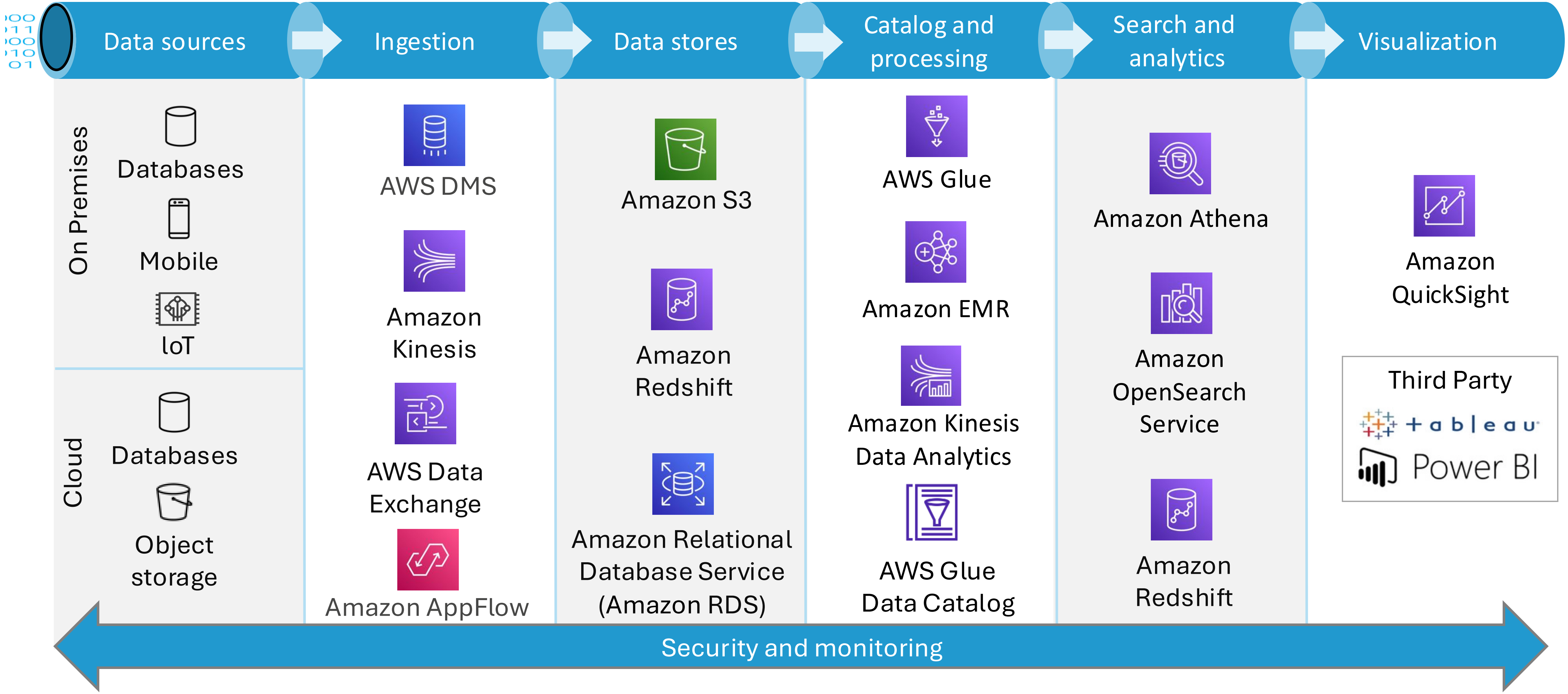
Comparison

Characteristics	Data warehouse	Data lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, websites, mobile apps, social media, and corporate application
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price and performance	Fastest query results using higher-cost storage	Query results getting faster using low-cost storage
Data quality	Highly curated data that serves as the central version of the truth	Any data that might or might not be curated (for example, raw data)
Users	Business analysts	Data scientists, data developers, and business analysts (using curated data)
Analytics	Batch reporting, BI, and visualizations	Machine learning, predictive analytics, data discovery and profiling

Analytics functionality



AWS services build the data lake



Data sources

On Premises

Databases

Mobile

IoT

Cloud

Databases

Object storage

Ingestion

AWS DMS

Amazon Kinesis

AWS Data Exchange

Amazon AppFlow

Data stores

Amazon S3

Amazon Redshift

Amazon Relational Database Service (Amazon RDS)

Catalog and processing

AWS Glue

Amazon EMR

Amazon Kinesis Data Analytics

AWS Glue Data Catalog

Search and analytics

Amazon Athena

Amazon OpenSearch Service

Amazon Redshift

Visualization

Amazon QuickSight

Third Party

tableau

Power BI

Security and monitoring

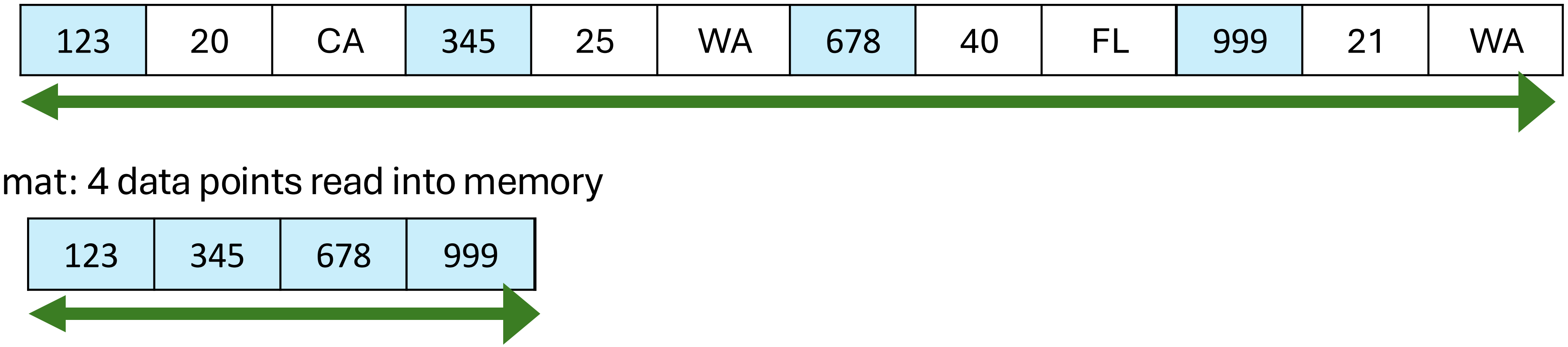
Row format compared to column format

Less data is read in columnar data

Example: Return all IDs in the dataset

ID	Age	State
123	20	CA
345	25	WA
678	40	FL
999	21	WA

Row format: 12 data points read into memory



123	20	CA	345	25	WA	678	40	FL	999	21	WA

Column format: 4 data points read into memory

123	345	678	999

Formatting: Columnar storage formats

Apache Parquet and ORC – Columnar storage formats optimized for fast retrieval of data.

Sample query: select l_orderkey from lineitem where l_partkey = 17766770	Data Scanned (GB)	Run time (seconds)
Text GZIP data	22	33.06
Parquet GZIP data with no sorting	2	1.72
Parquet GZIP data sorted on l_partkey	.034	1.0