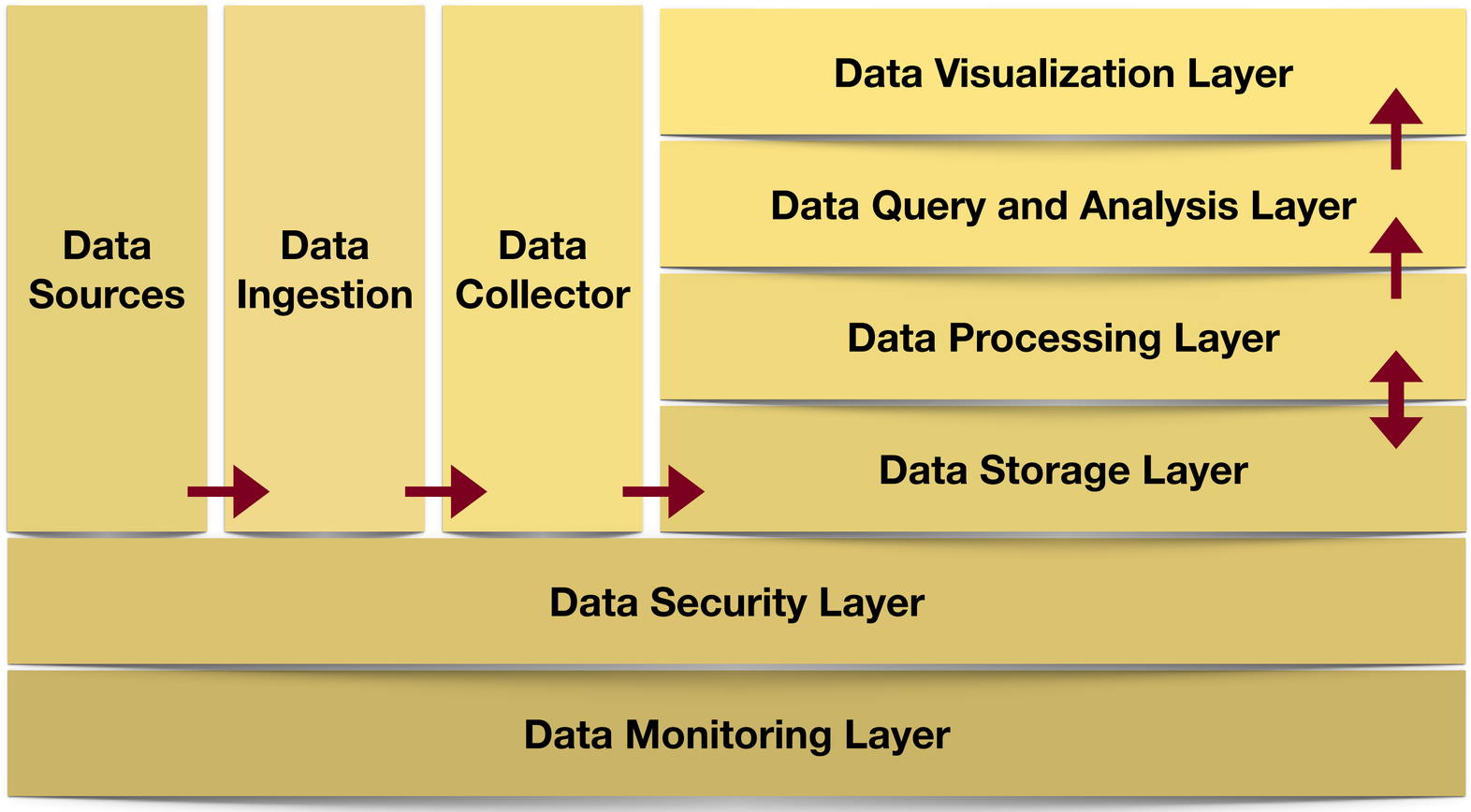
Big Data Architecture



Ingestion Layer

- Ingestion: the process of absorbing information.
- Data ingestion gathers data from data sources, prioritizes, validates, and routes it to the next layer.
- Data sources are:
  - Relational DBMS
  - Social networks
  - Emails, web pages
  - Multimedia data

Collector Layer

- Transports data from the data ingestion layer to the next layer
- Decouples the components to allow analytical processing

Storage Layer

- Critical layer to big data, since it is all about data.
- The volume and velocity directly impacts this layer.
- The storage solution should be appropriate for the data ingestion requirement
- Tools:
  - HDFS: Hadoop Distributed File System
  - GlusterFS: Dependable Distributed File System
  - Amazon S3 Storage Service

Processing Layer

- The layer where the analytical process begins.
- The needed data and features are selected, cleaned and processed.
- The processing styles:
  - Batch
  - Real time
  - Hybrid

Query Layer

- Active analytical processing takes place.
  - Apache Hive
  - Apache Spark SQL
  - Amazon Redshift
  - Presto - SQL query engine for Big Data
  - And so on

Visualization Layer

- Users perceives the value of data and analysis of data.
- Approaches:
  - Custom Dashboards
  - Real-Time Visualization Dashboards
  - Data visualization with Tableau
  - Exploring data sets with Kibana
  - Intelligence Agents
  - Recommendation Systems

Security Layer

- The crucial part of any sort of data.
- The essential aspect of Big Data Architecture.
- It is implemented at all layers.
  - Data authentication
  - Access control
  - Encryption and data masking
  - Auditing data access

Monitoring Layer

- The process of monitoring, auditing, testing, managing, and controlling the data.
- Continuous monitoring of data is an important part of the data governance mechanisms.
  - Data profiling and lineage
  - Data quality
  - Data cleansing
  - Data loss and prevention

Top Six Best Practice

- Be clear about the objectives.
- Authorize files access with predefined security policy.
- Implement Testing in Big Data.
- Implementing Big Data in a business decision.
- Safeguard sensitive encryption of information while at rest.
- Use agile solutions.

Congratulations

- Now you finished Module 3: Big Data.
- Now you have a better understanding about the big data around us.
- Structured data, data warehouse, and big data, are addressing different data needs.
- It is true that data is the future.
- See you soon in Module 4!