cloudera
Ask Bigger Questions

Apache Hadoop in the Enterprise

Dr. Amr Awadallah, CTO/Founder
@awadallah, aaa@cloudera.com

Cloudera

The Leader in Big Data Management Powered by Apache Hadoop™

The Leading Open Source Distribution of Apache Hadoop

Powerful Suite of System & Data Management Software

Built for the Enterprise

Founded: 2008

Employees: 450+

Customers: Over **50% of the Fortune 50** and **65% of the Fortune 500** plus top US intelligence and defense agencies

Partner Ecosystem: 700+ in hardware, software, and services

Education: 15,000+ trained annually; developers, admins, analysts, data scientists

Community: Founders and top supporters of the Hadoop open source ecosystem

cloudera
Ask Bigger Questions

Cloudera's Mission

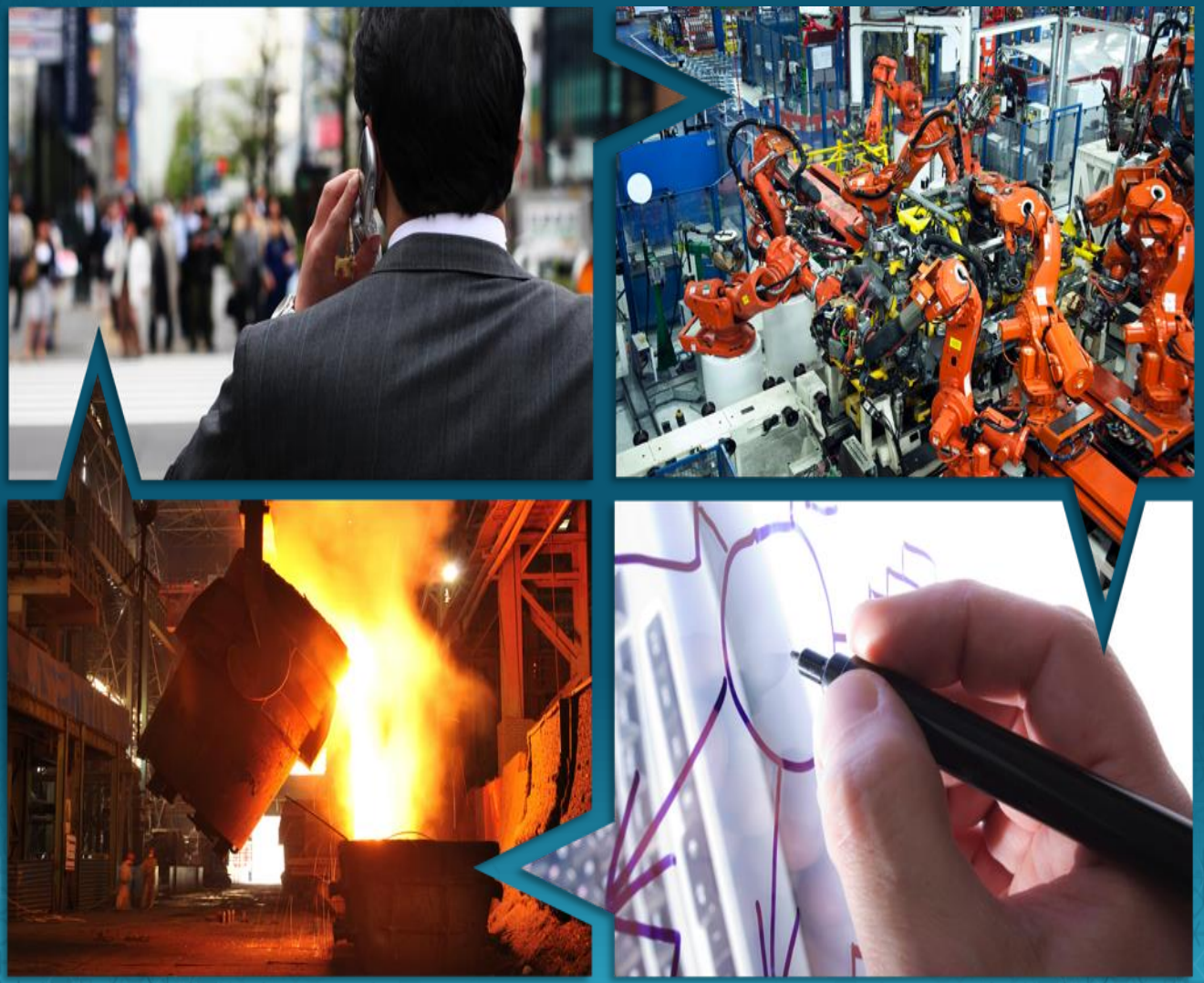
Help Organizations Gain Value from **All** Their Data

Solve data problems.

Solve problems with data.

Ask Bigger Questions.

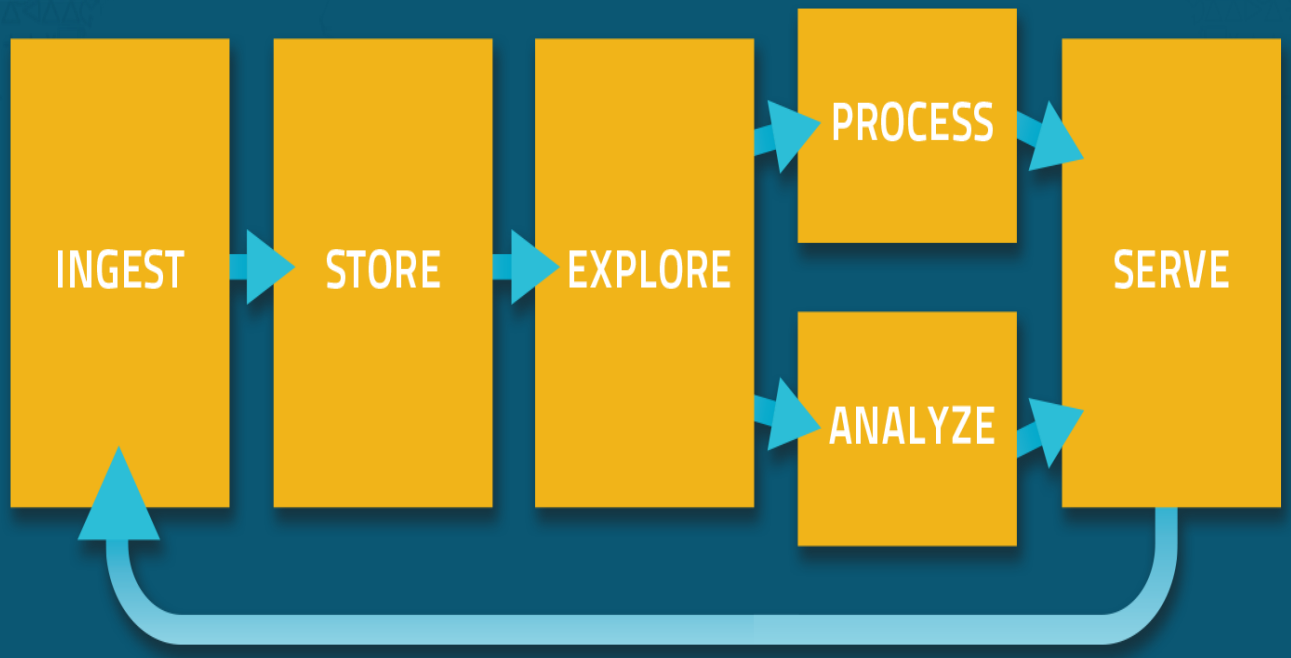
Why is This Happening Now?



10TB to 10PB

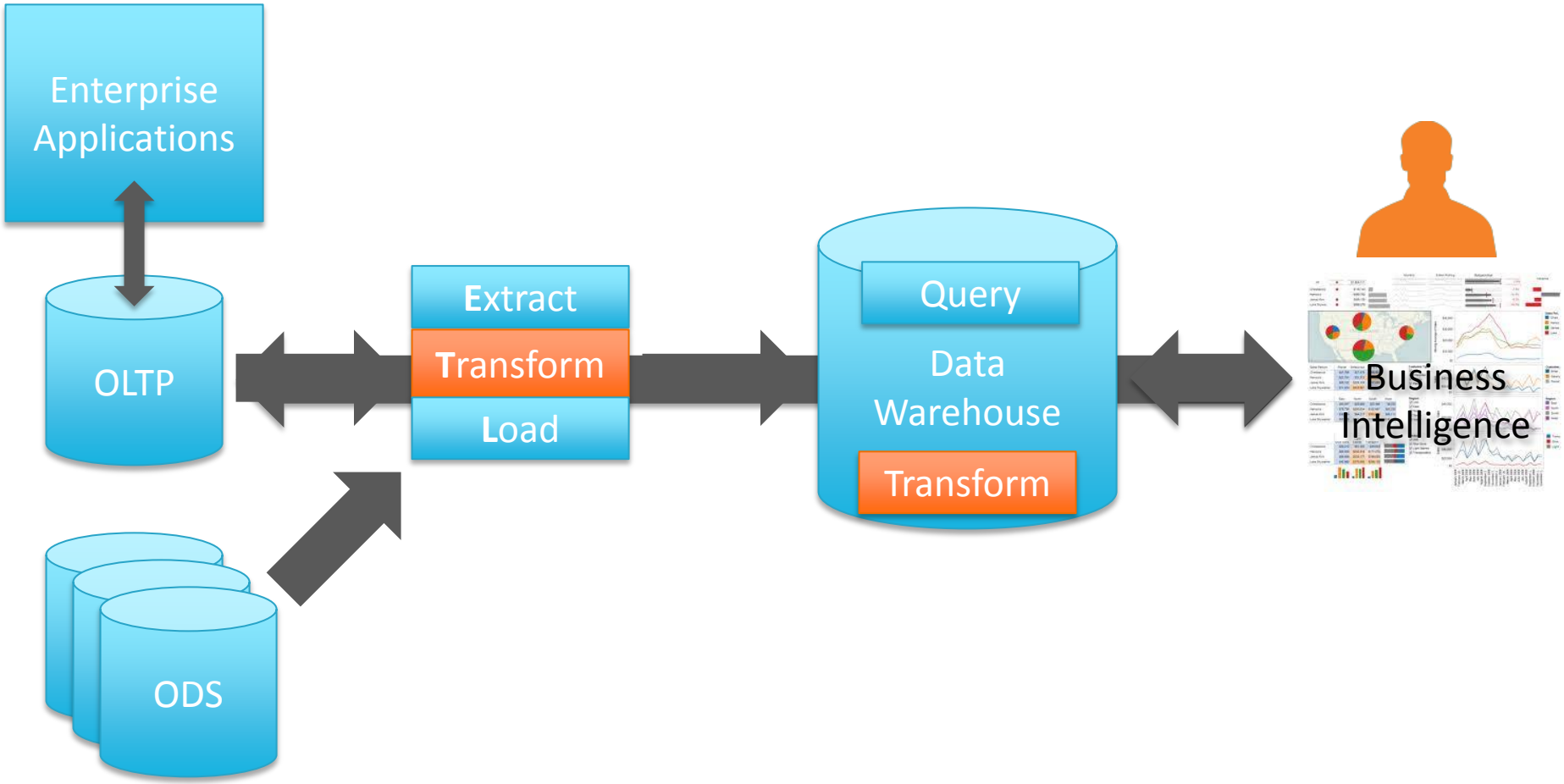
IT'S ALL (BIG) DATA (NOT)

Complications of Status Quo

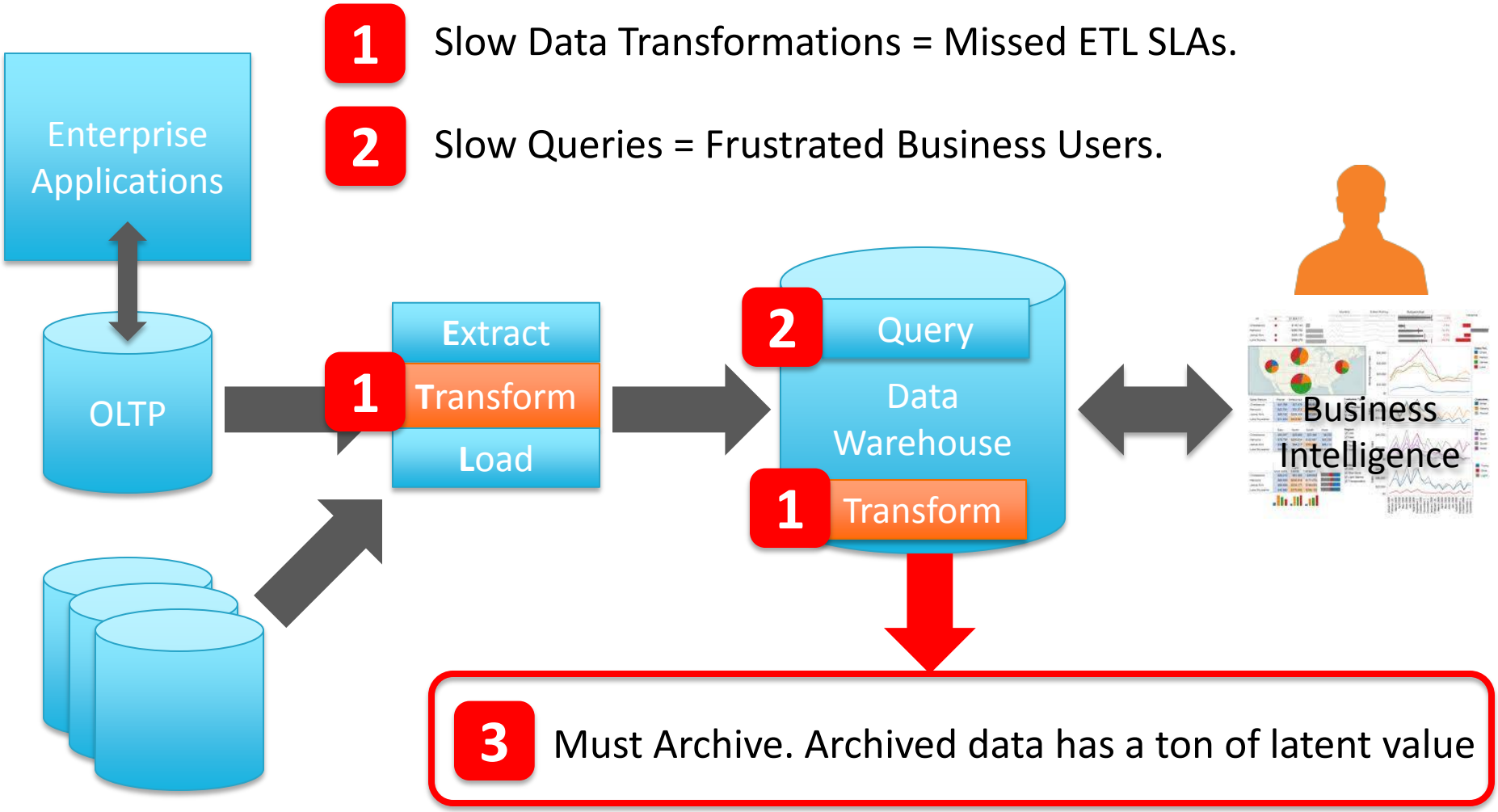


Structure Storage Network Silos

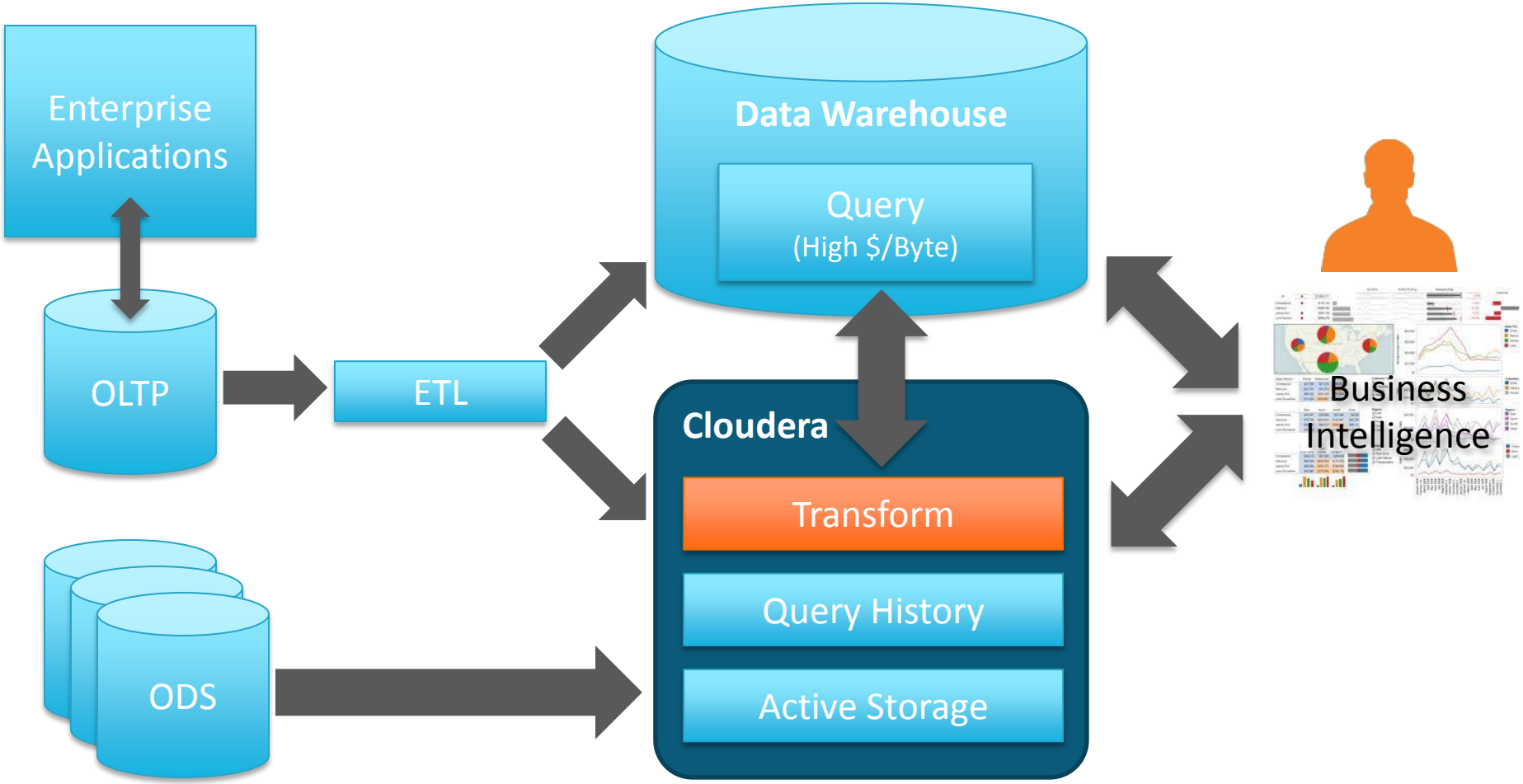
The Story of "T"



Volume, Velocity, Variety = Problems



Data Warehouse Optimization



Our Vision: The Android of Big Data

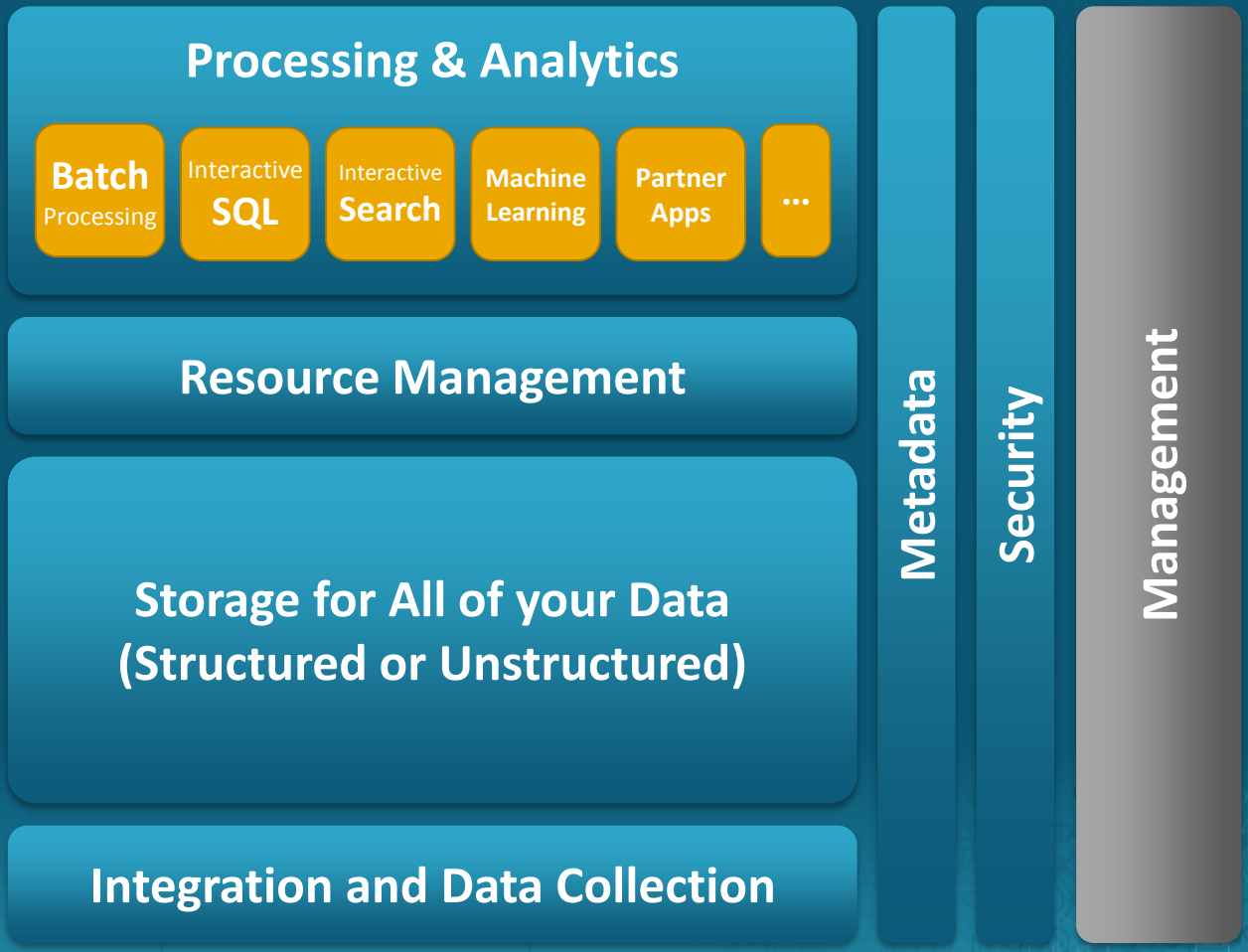
Scalable

Flexible

Cost-Effective

Open

Integrated



Cloudera Enterprise | The Platform for Big Data

Agility/Flexibility

Schema-on-Write (RDBMS):

- **Prescriptive Data Modeling:**
  - Create static DB schema
  - Transform data into RDBMS
  - Query data in RDBMS format
- New columns must be added explicitly before new data can propagate into the system.
- **Good for Known Unknowns (Repetition)**

Schema-on-Read (Hadoop):

- **Descriptive Data Modeling:**
  - Copy data in its native format
  - Create schema + parser
  - Query Data in its native format (does ETL on the fly)
- New data can start flowing any time and will appear retroactively once the schema/parser properly describes it.
- **Good for Unknown Unknowns (Exploration)**

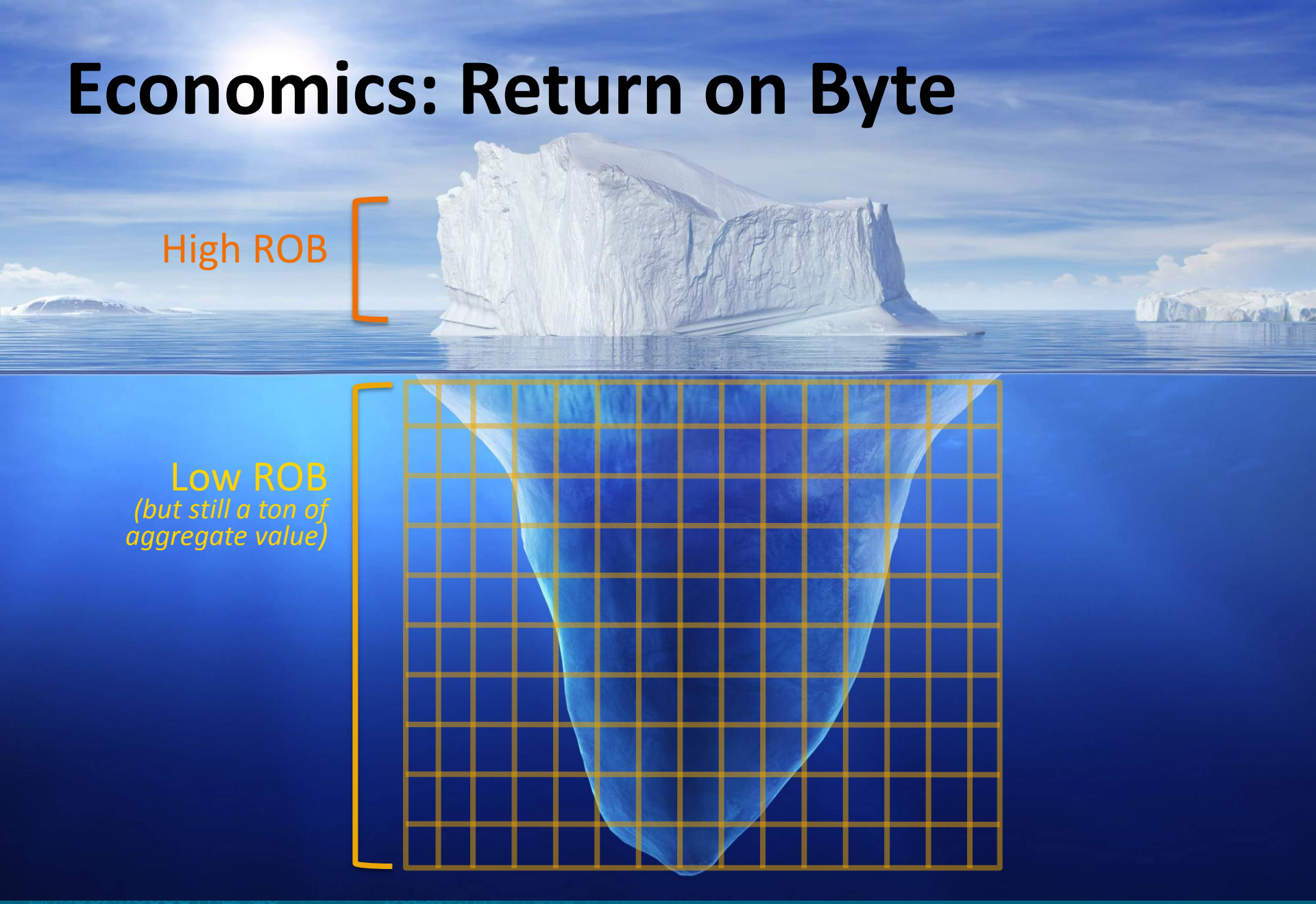
Scalable Technology + Scalable Development

Grows without requiring developers to re-architect their algorithms/application

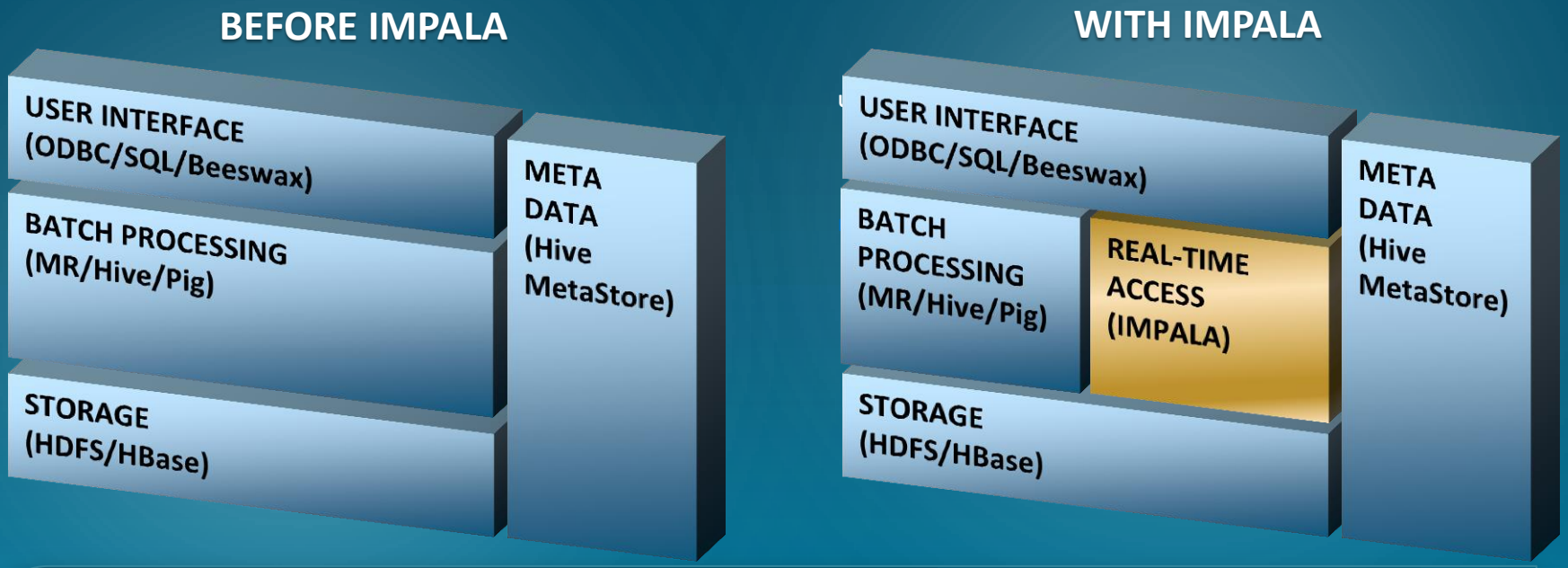
AUTO SCALE

cloudera
Ask Bigger Questions

Economics: Return on Byte



Cloudera Impala



- **Unified storage:**
  - Supports HDFS and HBase
  - Flexible file formats and schemas
- **Unified Metastore**
- **Unified Security**
- **Unified Client Interfaces:**
  - ODBC/JDBC
  - SQL syntax
  - Hue Beeswax Web UI

- **With Impala:**
  - Interactive ANSI-92 SQL queries
  - Native distributed query engine
  - Optimized for low-latency
- **Provides:**
  - Answers as fast as you can ask
  - Everyone can ask questions of all data
  - Big data storage and analytics together

But What about the RDBMS?

"Use right tool for the right job"

Optimize existing EDW systems for
high-performance operational analytics

KEEP IN EDW

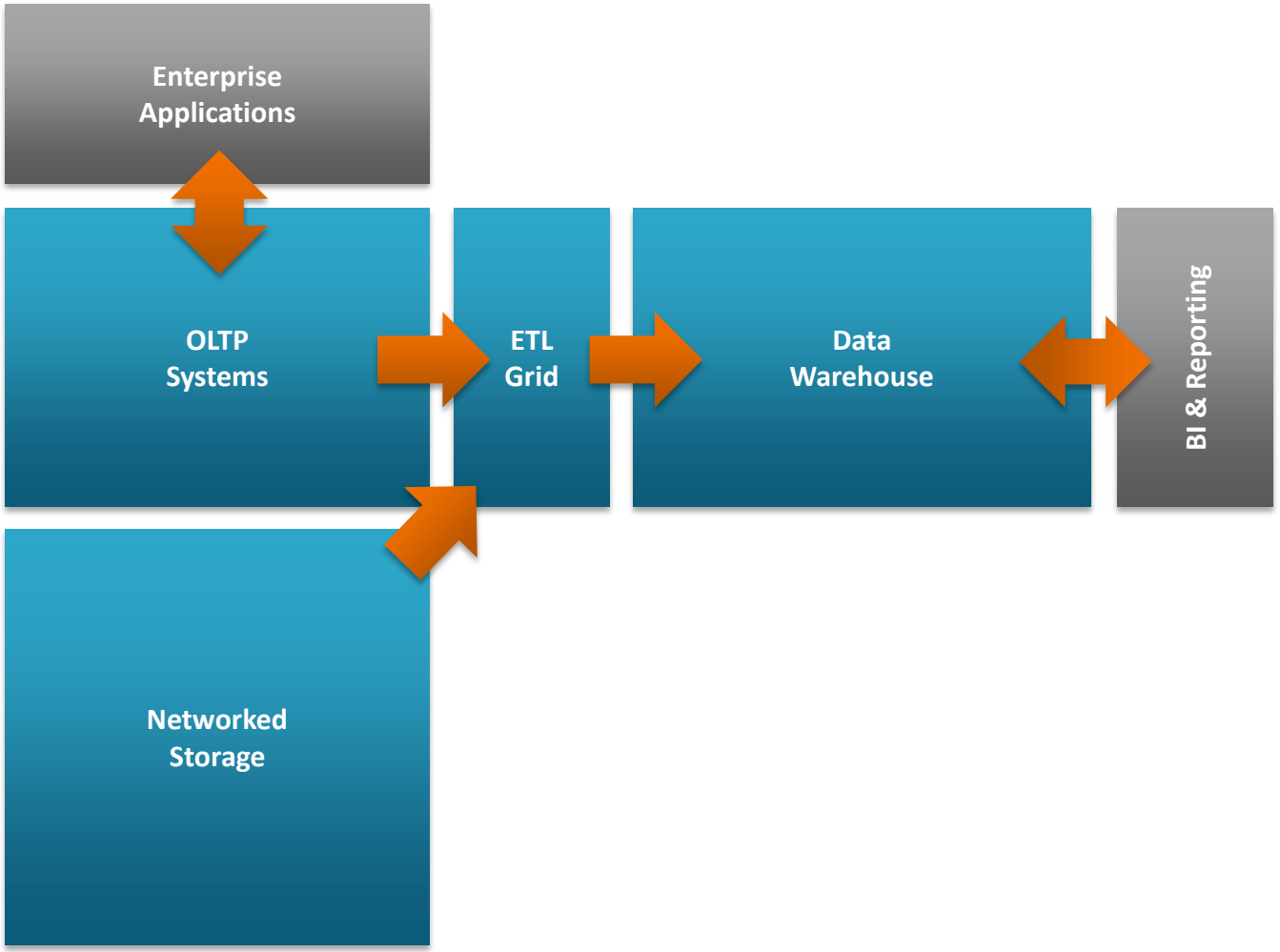
- Operational Analytics
- Reporting
- Multi-statement Transactions

MOVE TO CLOUDERA

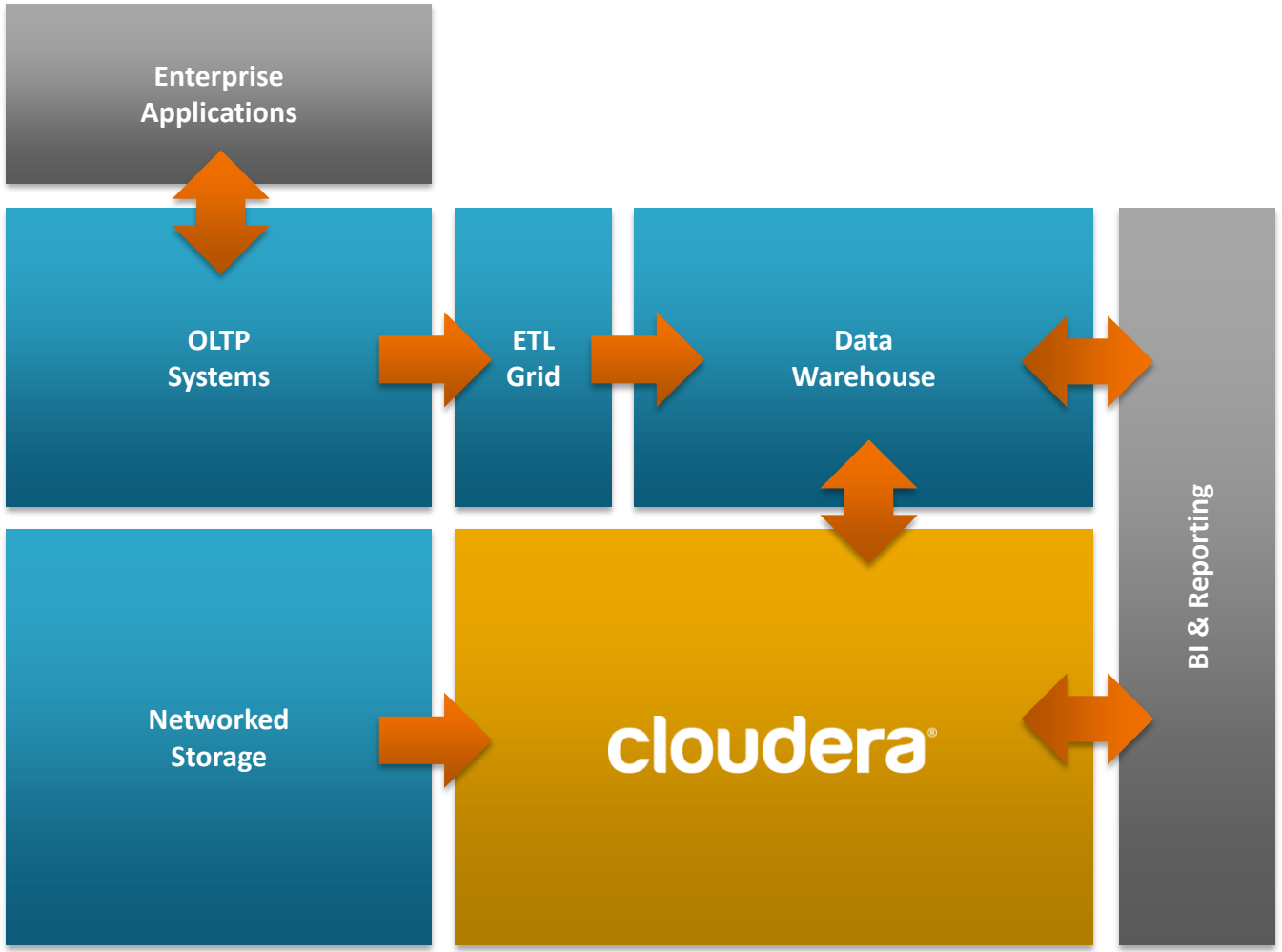
- Historical Data
- Data Processing
- Ad Hoc Exploration
- Transformation/Batch

cloudera
Ask Bigger Questions

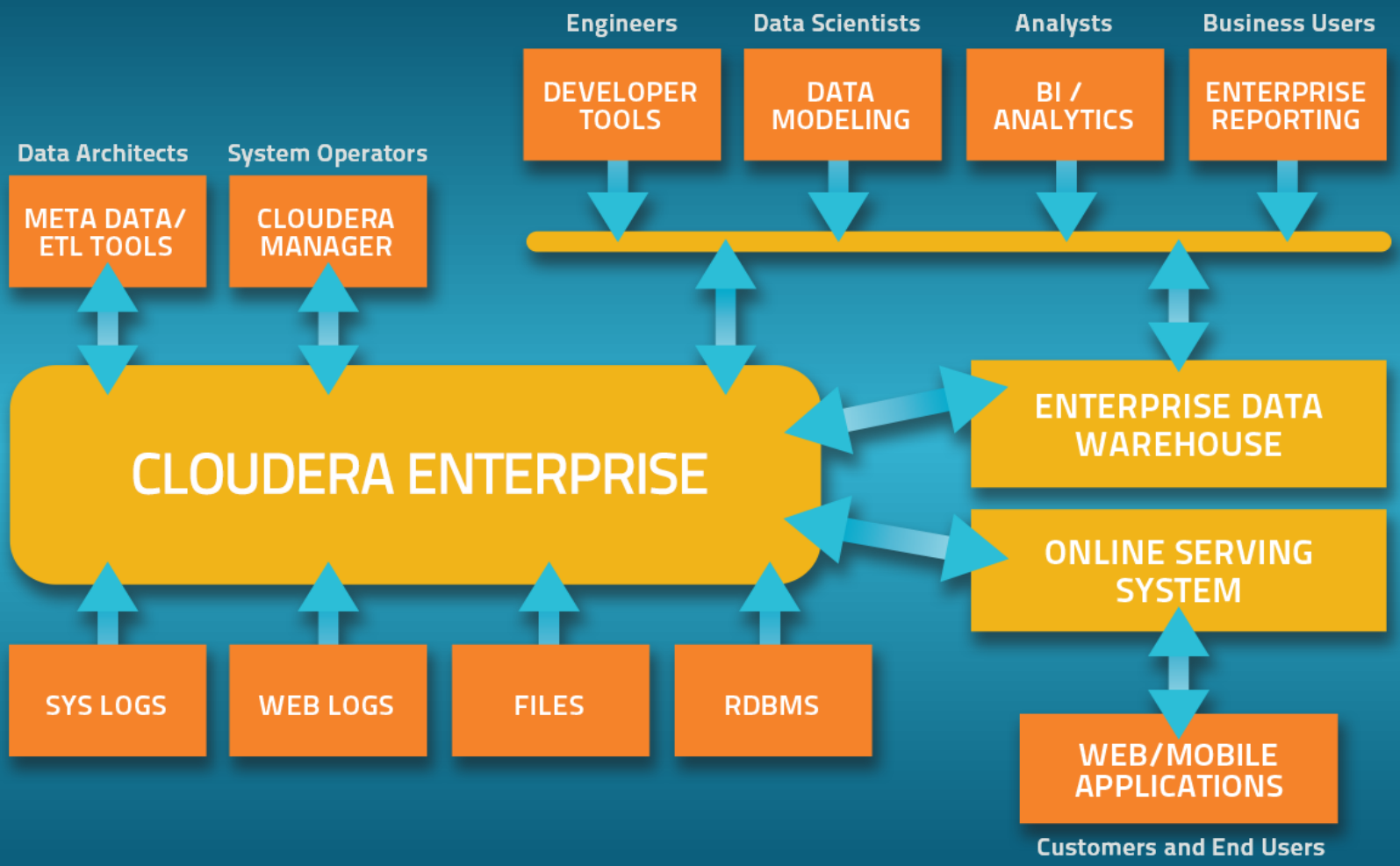
Legacy Information Architecture



New Information Architecture

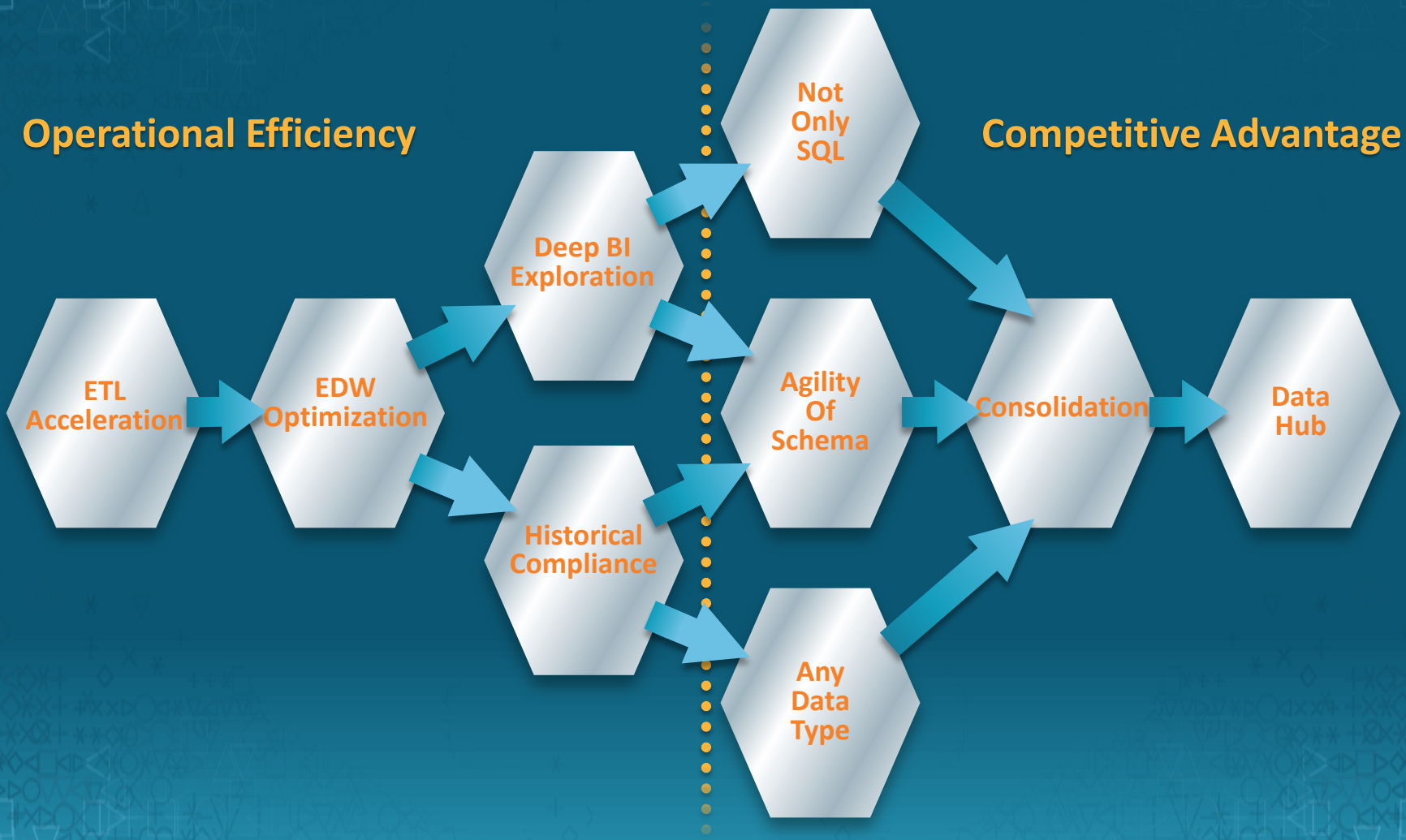


The New Enterprise Big Data Stack



cloudera
Ask Bigger Questions

Maturity Path



Beyond Data Warehousing

AUTOMOTIVE
Auto sensors reporting location, problems

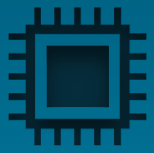
COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, customer service

FINANCIAL SERVICES
Risk & portfolio analysis
New products

EDUCATION & RESEARCH
Experiment
sensor analysis

HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg quality
Warranty analysis



LIFE SCIENCES
Clinical trials
Genomics

MEDIA / ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching
Website optimization

HEALTH CARE
Patient sensors, monitoring, EHRs Quality of care

OIL & GAS
Drilling exploration sensor analysis

RETAIL
Consumer sentiment
Optimized marketing

TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows
Customer sentiment

UTILITIES
Smart Meter analysis for network capacity

LAW ENFORCEMENT & DEFENSE
Threat analysis,
Social media monitoring,
Photo analysis

cloudera
Ask Bigger Questions

The Cloudera Platform for Big Data

Key Use Cases:

- Transformation Offload (aka **ETL/ELT Offload**)
- Exploratory Archive (aka **Active Archive**)

Benefit 1: Flexibility

- Store any data
- Run any analysis
- Keep's pace with the **rate of change** of incoming data

Benefit 2: Scalability

- Proven growth to PBS/1,000s of nodes
- No need to rewrite queries, automatically scales
- Keep's pace with the **rate of growth** of incoming data

Benefit 3: Economics

- Cost per TB at a fraction of other options
- Keep all of your data alive in an active archive
- Powering the data beats algorithm movement

Dr. Amr Awadallah
CTO/Founder
@awadallah
aaa@cloudera.com