cloudera

cloudera

New Software Trends

Hadoop and Related Software

Jeff Hammerbacher
Chief Scientist, Cloudera
May 22, 2012

Presentation Outline

- 1. Philosophy
- 2. Platform
- 3. Recent Developments

1. Philosophy

Philosophy

- The true challenges in the task of data mining
  - Creating a data set with the relevant and accurate information
  - Determining the appropriate analysis techniques

Adapted from “Exploratory Data Mining and Data Cleaning” by Tamraparni Dasu and Ted Johnson

Philosophy

Creating a data set

- Store all of your data in one place
- Data first, questions later
- Store first, structure later
- Keep raw data forever

Philosophy

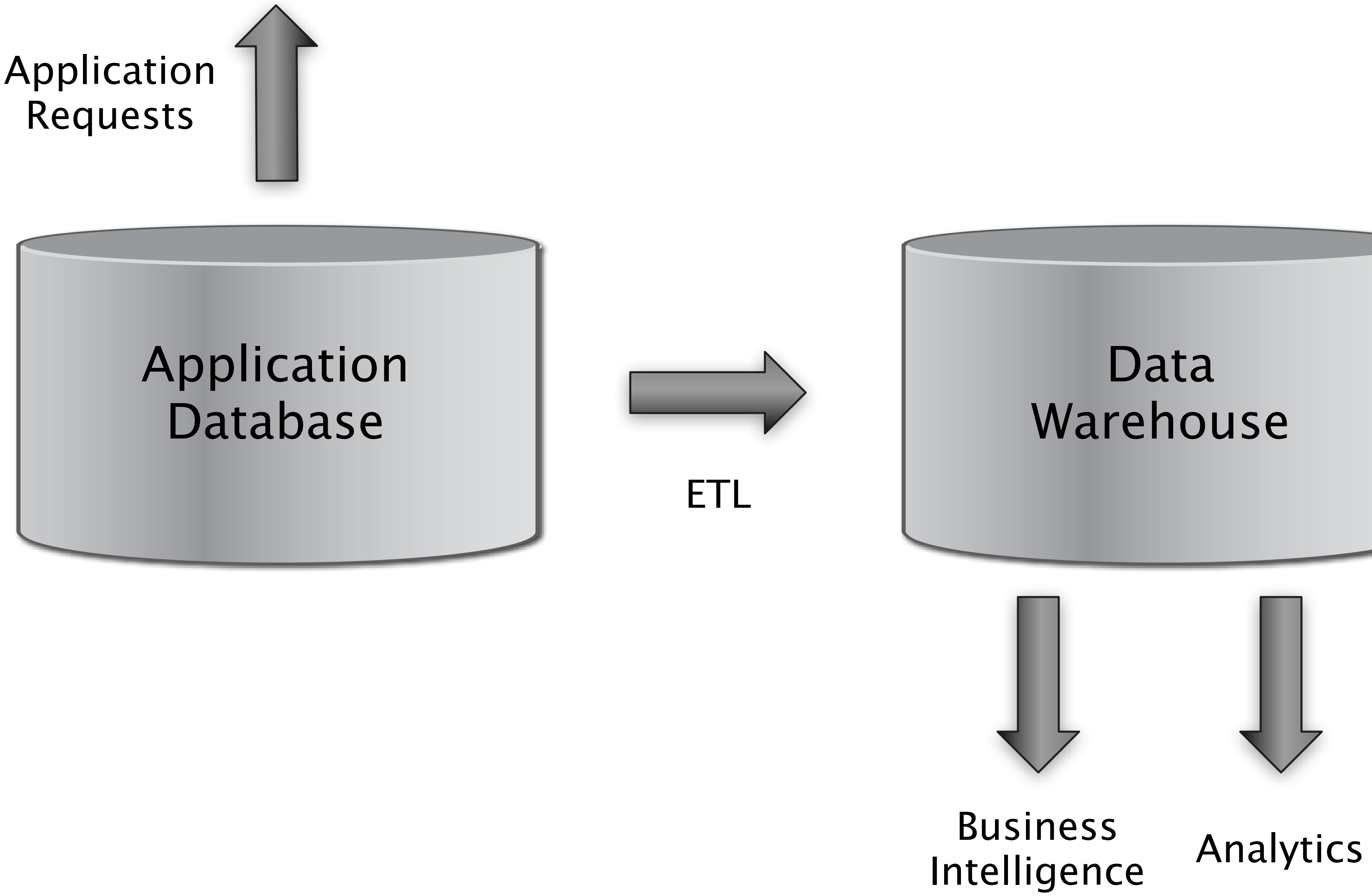
Choosing an analysis technique

- Enable everyone to party on the data
  - Developers
  - Analysts
  - Business users

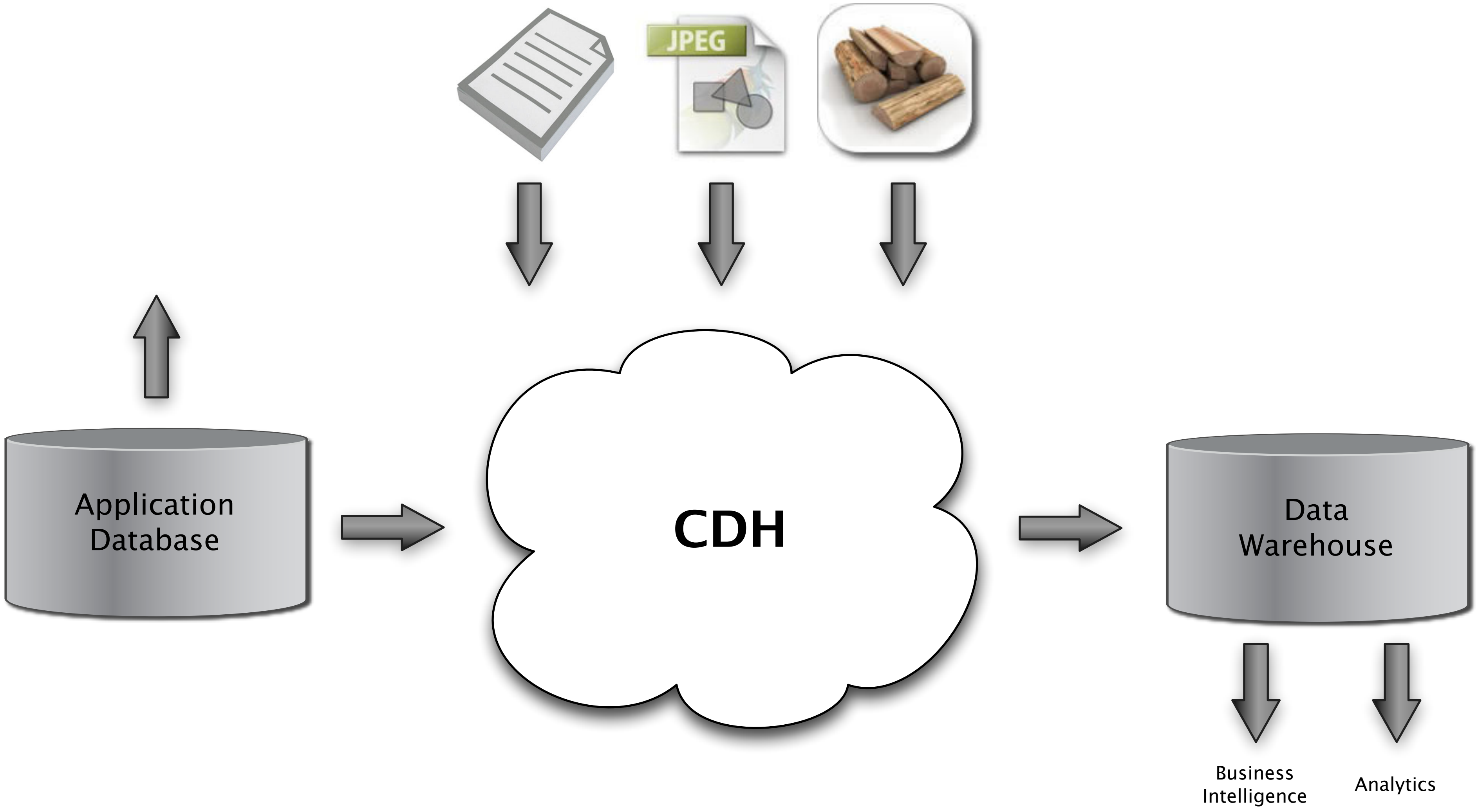
Philosophy

- We have to produce tools to support the whole research cycle
  - Data capture
  - Data curation
  - Data analysis
  - Data visualization

Adapted from “The Fourth Paradigm” by Jim Gray


Application
Requests
Application
Database
ETL
Data
Warehouse
Business
Intelligence
Analytics

cloudera


JPEG
Application Database
CDH
Data Warehouse
Business Intelligence
Analytics

2. Platform

Platform

Substrate

- Commodity servers
  - Open Compute
- Open source operating system
  - Linux
- Open source configuration management
  - Puppet, Chef
- Coordination service
  - ZooKeeper

Platform

Storage

- Distributed schema-less storage
  - HDFS
- Append-only table storage and metadata
  - Hive
- Mutable table storage and metadata
  - HBase

Platform

Compute

- Cluster resource management
  - YARN
- Processing frameworks
  - MapReduce, MPI
- High-level interfaces
  - Crunch, PigLatin, HiveQL, Oozie
- Libraries
  - DataFu, Mahout, Giraph

Platform

Integration

- Data access
  - FUSE
  - ODBC/JDBC
- Data ingest
  - Sqoop
  - Flume
- User interface
  - Hue

Platform

Management

- Cloudera Manager
  - Service configuration, deployment, and management
  - Host, service, and activity monitoring
  - Event management and alerting
  - Log management

3. Recent Developments

Recent Development

Substrate

- Fat servers with fat pipes
  - Better tolerance of single disk failures
  - Support for multiple network interfaces
- Operating system support for isolation
  - LXC on Linux, Control Groups on Windows
- Local filesystem improvements
  - btrfs
- Dynamic changes to cluster membership in ZooKeeper

Recent Development

Storage

- Recently completed
  - High availability
  - Multiple namespaces
  - Rolling upgrades
  - Unified file format and compression
- Upcoming
  - Distributed snapshots
  - Cross-data center replication
  - Separation of namespace and block management

Recent Development

Compute

- Recently completed
  - Stabilization of YARN and MR2
  - Hamster: MPI on YARN
  - Crunch
- Upcoming
  - Isolation and workload management
  - Low latency job scheduling
  - Additional frameworks

Recent Development

Integration

- Recently completed
  - Flume NG
  - Sqoop 2

Recent Development

Management

- Recently completed
  - Host inspector
  - API
  - Automatic configuration
  - i18n
  - Multiple cluster support
- Upcoming
  - Improved audit capabilities
  - Disaster recovery

cloudera