Strata CONFERENCE + HADOOP WORLD

Oct. 28–30, 2013
NEW YORK, NY

#strataconf + #hw2013

Co-presented by O'REILLY cloudera

Hadoop & the Data Warehouse–
When to use which

TERADATA | Hortonworks

FEATURED presenters

Hortonworks

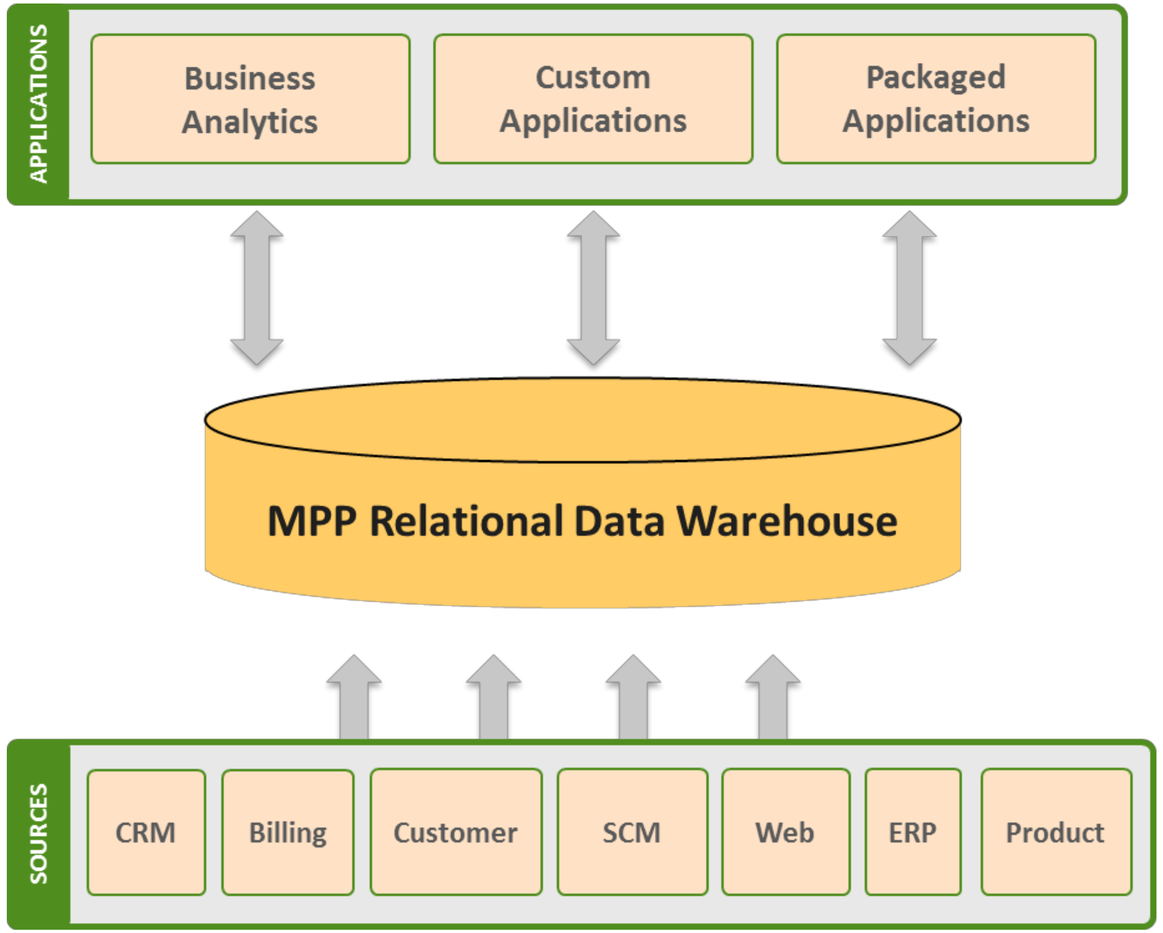
Ari Zilka
Chief Technology Officer, Hortonworks

TERADATA

Stephen Brobst
Chief Technology Officer, Teradata

MPP Relational Data Warehouse Design Core

What problems was this technology engineered to solve?



SQL Advantages with an MPP RDBMS

- Full ANSI SQL:
  - The *lingua franca* of business users when accessing data.
  - Decades of standardization (stable, feature rich, portable).
- Mature 3rd Party SQL based tools that provide business users with **self service** direct access to the data:
  - BI Tools.
    - In-database statistical packages.
    - Analytic applications (CRM, SCM, MDM).
- Easily parallelized.
- Scalable when manipulating large data sets.

ACID Advantages in an MPP RDBMS

- Guarantees database actions are processed reliably.
- Ensures 100% query result accuracy.
- Supports updates and deletes.
- Needed for applications that require 100% consistency.

Atomicity - All of the pieces are committed or none are committed.

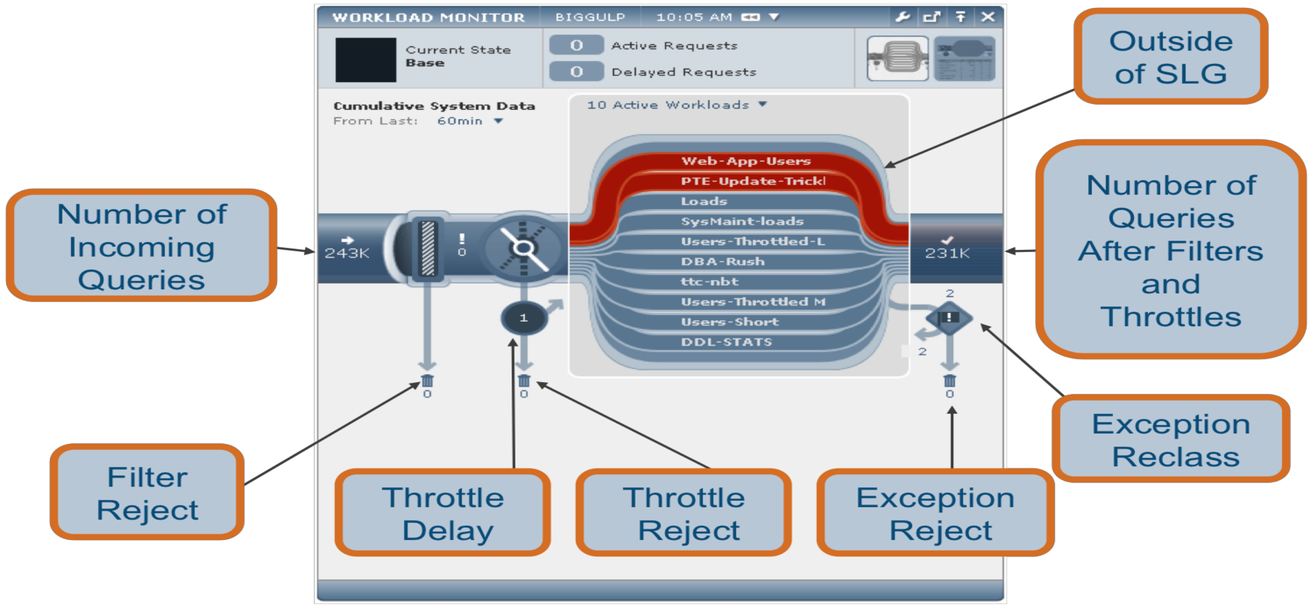
Consistency - Creates a new and valid state of data, or, if any failure occurs, returns all data to its original state.

Isolation - Processed and not yet committed transactions must remain isolated from any other transactions.

Durability - Committed data is saved such that in event of a failure and system restart, the data is available in its correct state.

Tight Vertical Integration

- End-to-end management of resources.
- Efficient utilization of resources.
- Engineered extremely well for known data.
- Fine-grained parallelism and resource management.
- Consistency of service level delivery.

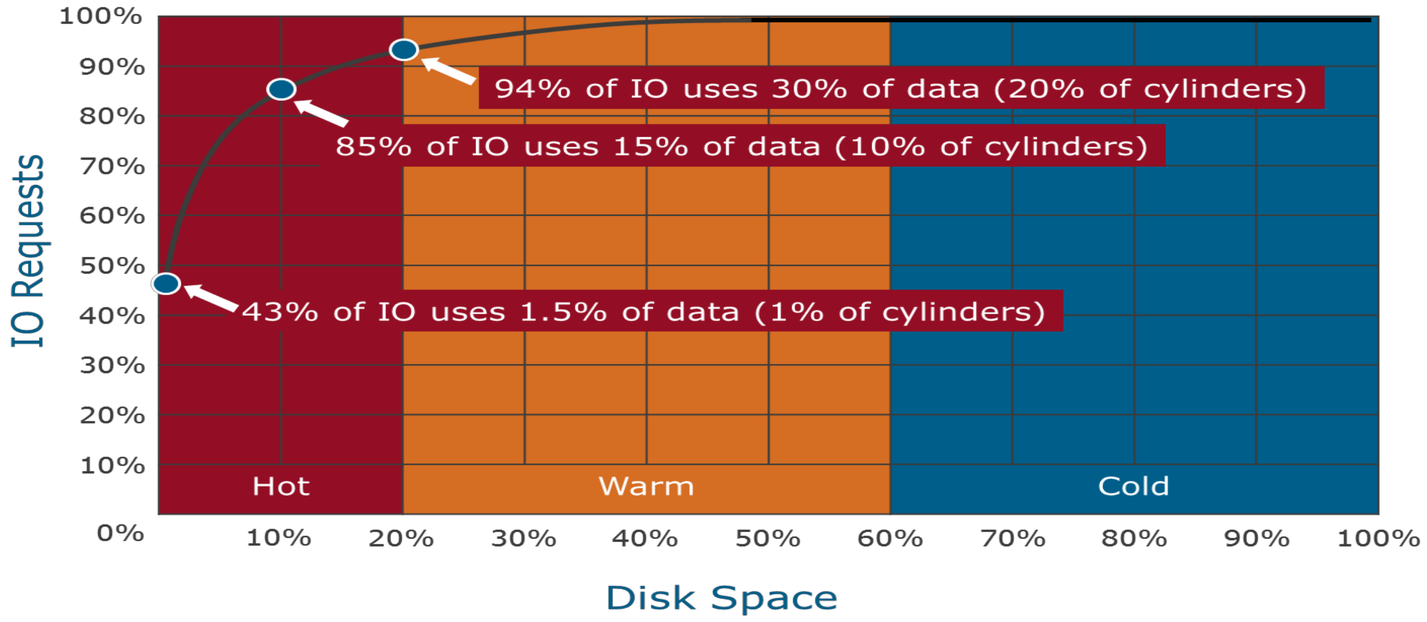


Best Practices Management:

- Workload functions .
- Workload groups.
- Exceptions.
- Priorities.
- Time periods.

Low Latency Advantages of MPP RDBMS

- Indexes.
- Statistics.
- Advanced partitioning.



Multi-temperature storage with automated distribution of data based on access patterns:

- In-Memory.
- Solid-State Drives.
- Fast Hard Drives.
- Fat Hard Drives.

Cost Based Optimizer Advantages in an MPP RDBMS

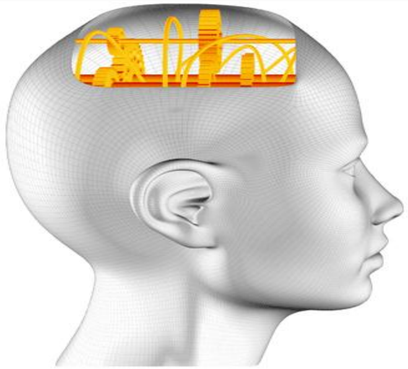
- Best practices optimizer determines how the query will be processed most efficiently, with no “hints” or degrees of parallelism necessary.

- In chess, you can look out a few moves to decide your best next move, but you can’t envision all move and countermove sequences for the entire game:
  - **The Grand Master** has the knowledge, experience, and intelligence to identify and use the right strategy.
  - With Hadoop, the **user** takes a heavy role in optimizing the execution of queries.
  - With an MPP RDBMS, the ***software*** is the optimizer.

Many ways to process a complex query…

Query Rewrite
- semantic optimization
- different types of vendor tools

Query Complexity
- Join costing & planning
- Aggregation



Fast/Efficient Data Access
- Access path - Indexing
- Partitioning (CP & PPI)
- Advanced partitioning schemes (range & case based, multilevel, dynamic)
- IO Optimizations (efficient scans/sync scan) scan optimization

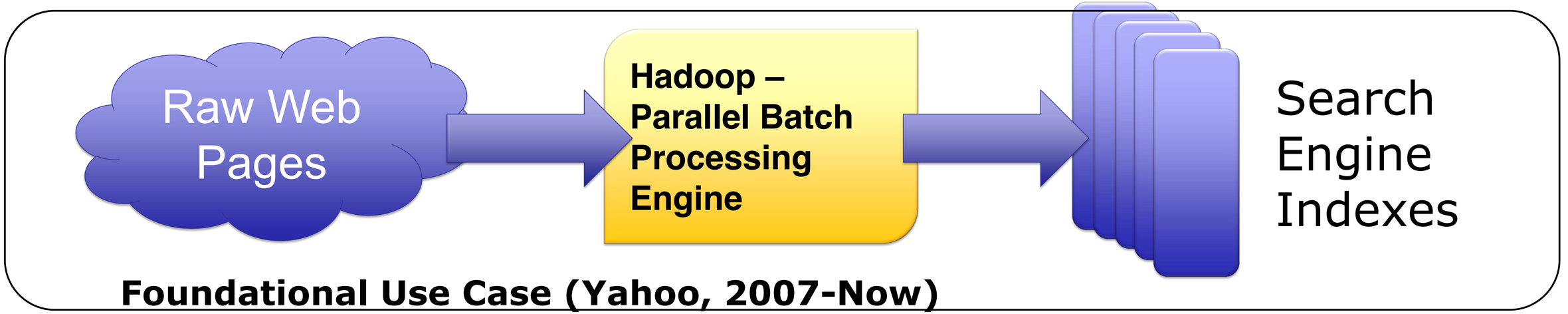
Granular Security Advantages in an MPP RDBMS

- Row level security.
- Column level security.

- An MPP RDBMS tightly integrates mature security features:
  - User-level security controls.
  - Increased user authentication options.
  - Support for security roles.
  - Enterprise directory integration.
  - Auditing and monitoring controls.
  - Encryption.

MPP RDBMS Customer Successes

Hadoop Design Core



- Started as part of open source search engine project Nutch in 2004; spun-off and adopted by Yahoo in 2006.
- What mattered to Yahoo in engineering Hadoop?
  - Scalability.
  - Ability to handle non-relational data ➔ web page text and their links.
  - Processing power & flexibility ➔ search index generation algorithms.
- What did *not* matter to Yahoo?
  - Hardware efficiency (abundance of servers).
  - Usability/manageability/security (abundance of engineers).
  - Immediate response time & latency (batch nature of problem).

Capture Non-Relational Types of Data

- Clickstream
  Capture and analyze website visitors' data trails and optimize your website.
- Sensor/Machine
  Discover patterns in data streaming automatically from remote sensors and machines.
- Geographic
  Analyze location-based data to manage operations where they occur.
- Server Logs
  Research logs to diagnose process failures and prevent security breaches.
- Text
  Data on how your customers feel about your brand and products.
- Rich format (video, pictures, etc.)
  Understand patterns in image files, satellite images, surveillance video, etc.

Value

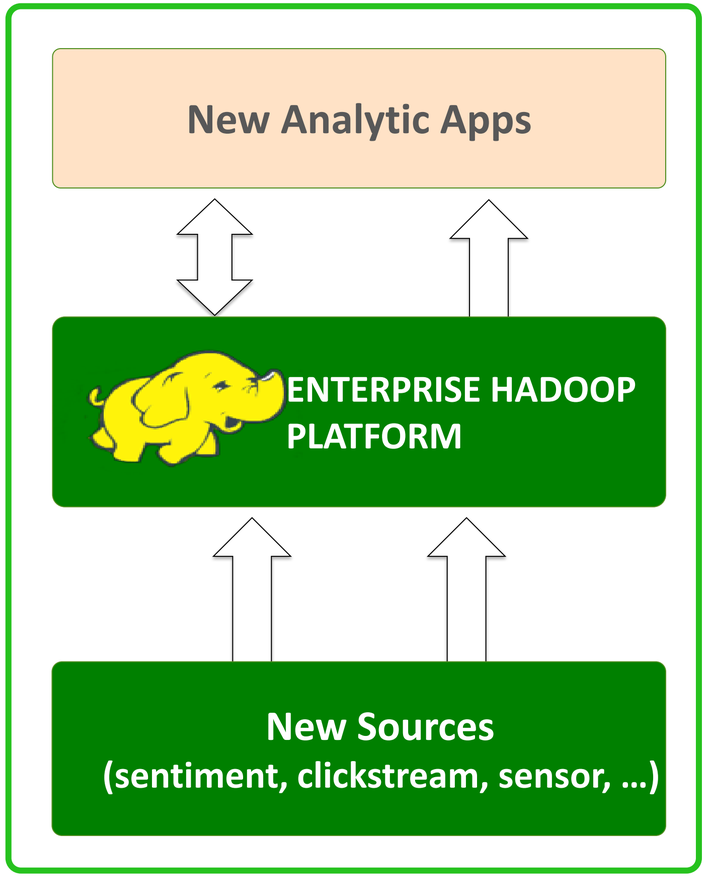
Complex, Flexible, Processing = New Analytic Apps

- Schema on read.

- No schema advantages enable easier combination of data.
  - Raw data format provides complete flexibility.
  - Non-traditional data types easily supported (graph, text, weblog, etc.).
  - NoETL approach provides agility.
  - Late-binding gives more power to the data scientist.

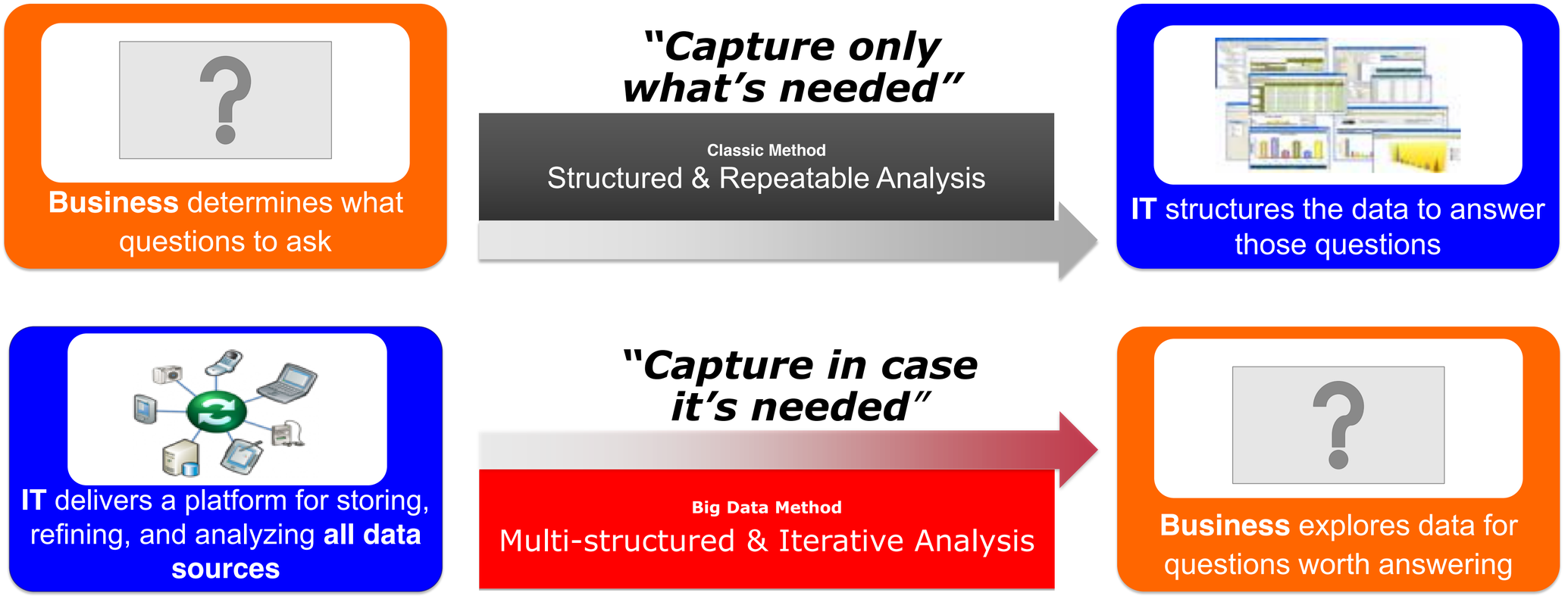
- Flexible programming language choices.

- Not constrained to SQL processing model.



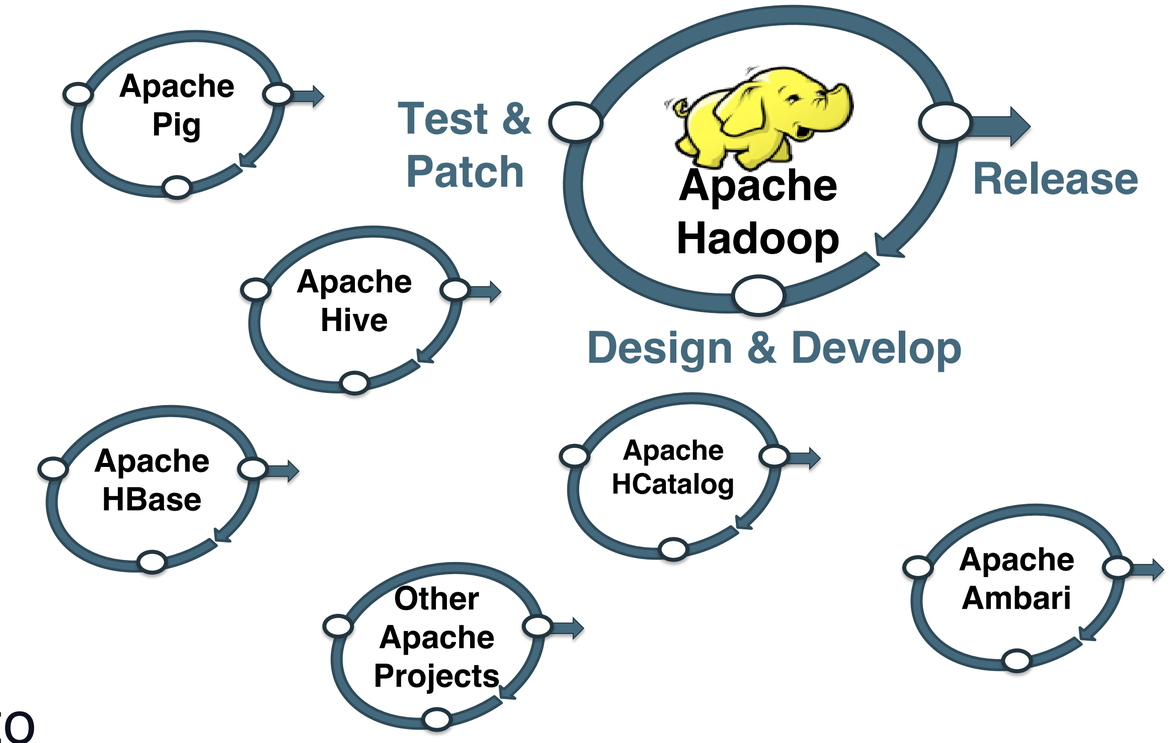
Use Hadoop to Create a Share Data Lake

- Makes keep “all” data “forever” financially viable.
- Better ROI for finding golden needles in very large data haystacks.
- Evolution is from a single analytic application to a data lake.



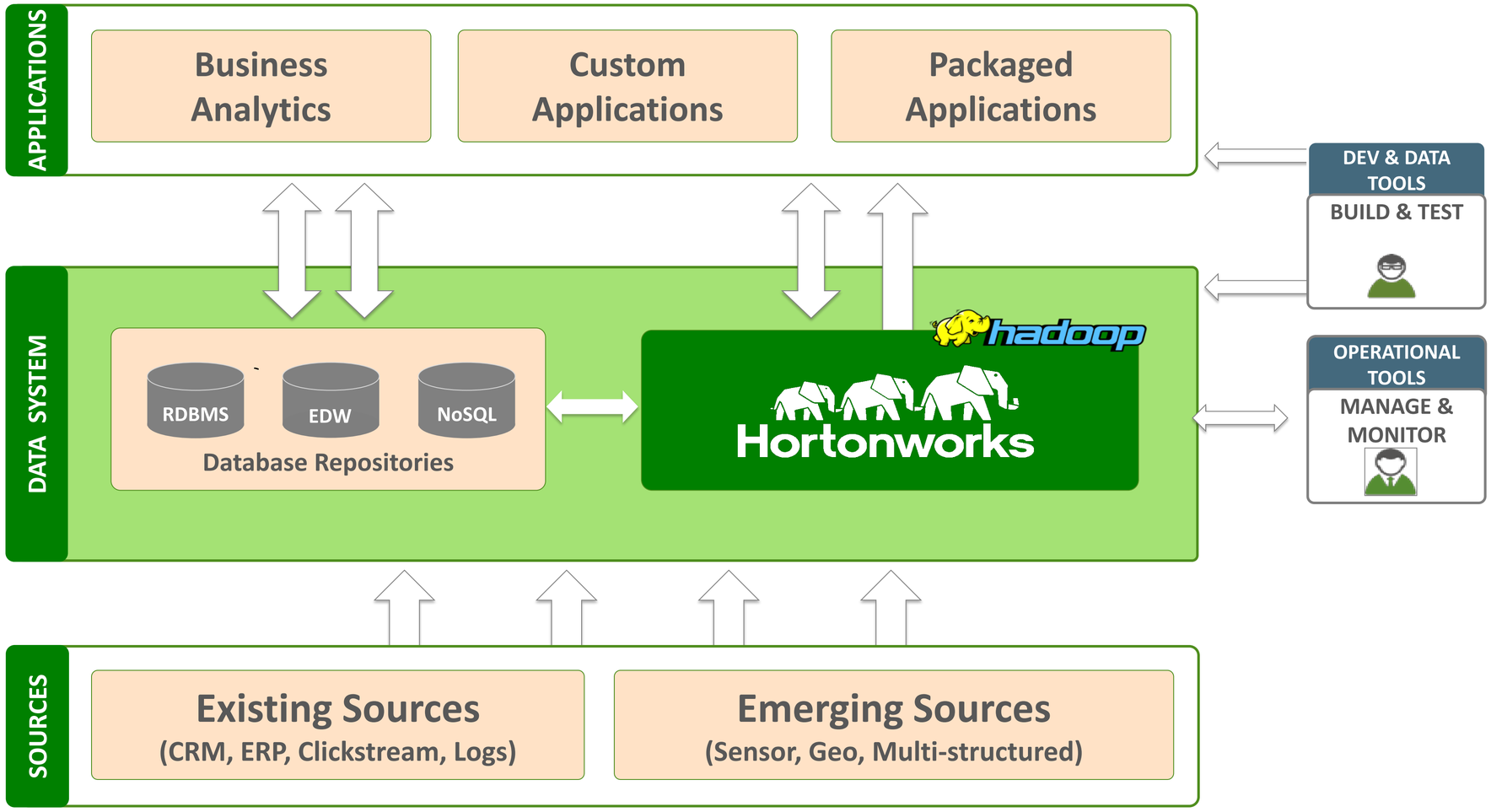
Open Source Community Advantages for Hadoop

- High degree of innovation.
- Open and transparent processes.
- Multiple contributors.
- Broad, creative thinking.
- Different perspectives.
- Define projects and then work with community to address.
- Commit everything back into the community, with nothing held back.
- No proprietary software; contribute all to Apache. No vendor lock in.



Hadoop Customer Successes

Next Generation Data Architecture Enabled



Key Considerations

MPP RDBMS		Hadoop
Stable Schema	⟷	Evolving Schema
Leverages Structured Data	⟷	Structure Agnostic
ANSI SQL	⟷	Flexible Programming
Iterative Analysis	⟷	Batch Analysis
Fine Grain Security	⟷	Coarse Grain Security
Cleansed Data	⟷	Raw Data
Seeks	⟷	Scans
Updates/Deletes	⟷	Ingest
Service Level Agreements	⟷	Flexibility
Core Data	⟷	All Data
Complex Joins	⟷	Complex Processing
Efficient Use of CPU/IO	⟷	Low Cost of Storage

Thank You!

Questions and Answers