上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Big Data, Big Challenge

—From Hadoop Perspective

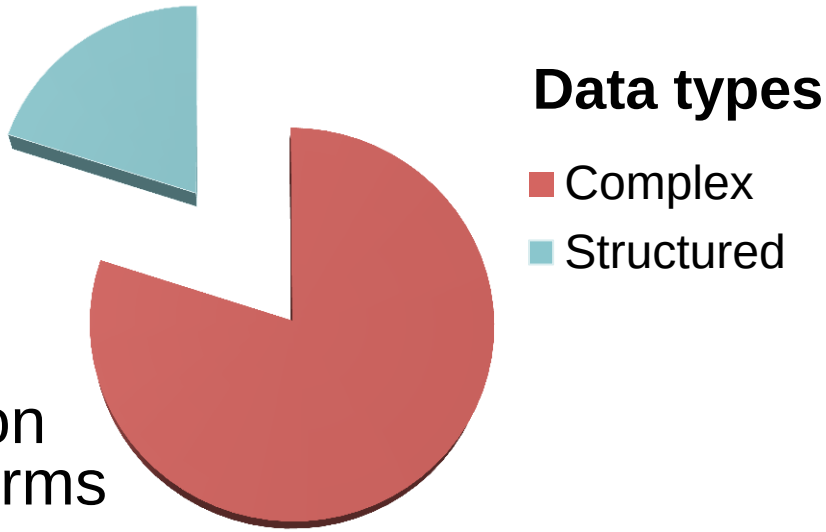
Hao ZHU

REINS LAB

Complex Data is Growing Really Fast

Gartner – 2009

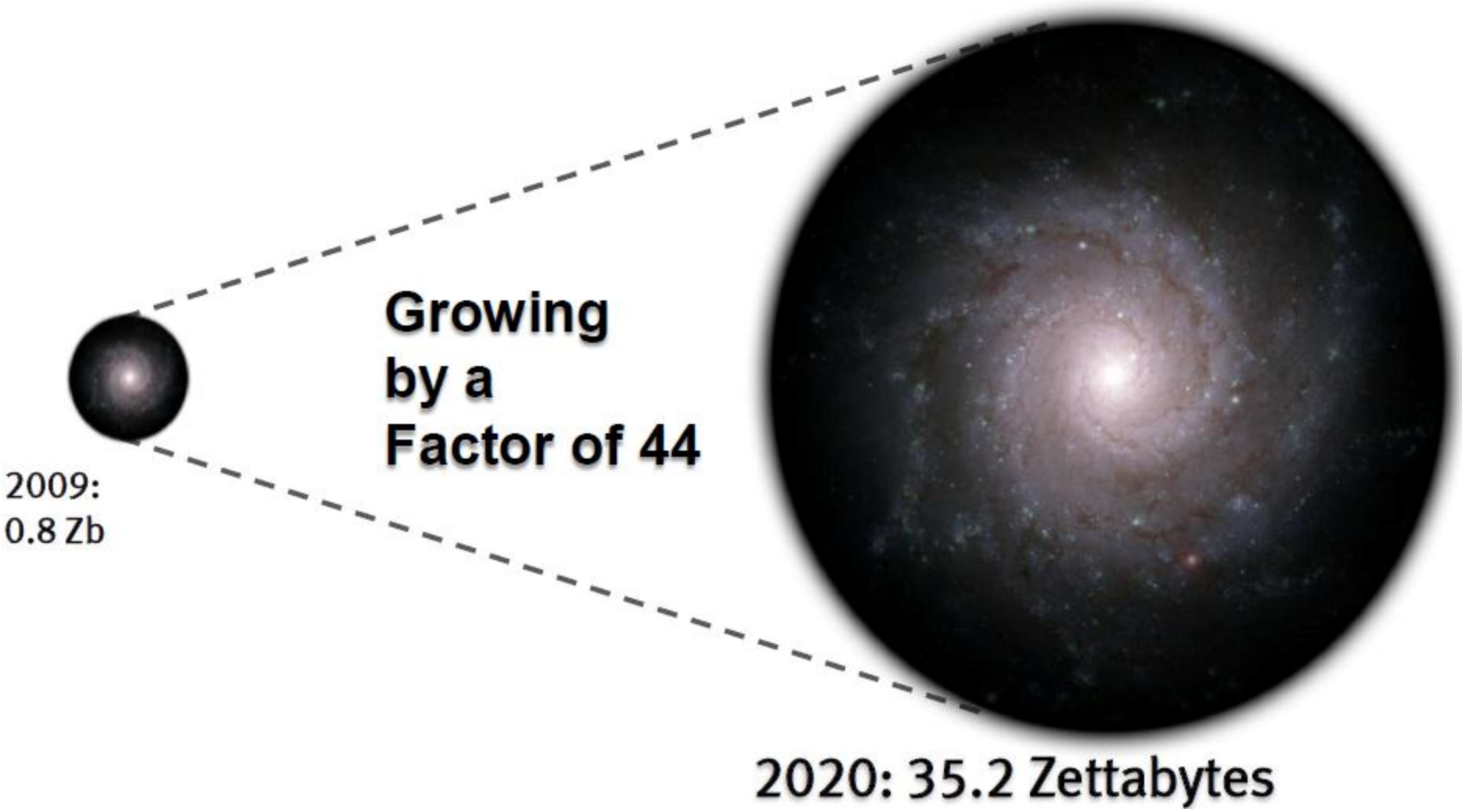
- Enterprise Data will grow 650% in the next 5 years.
- 80% of this data will be unstructured (complex) data



IDC – 2008

- 85% of all corporate information is in unstructured (complex) forms
- Growth of unstructured data (61.7% CAGR) will far outpace that of transactional data

The Digital Universe 2009-2020



上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Big Data is Changing the World

Expanding Data Sources

Science and research
- Gene sequences
- LHC accelerator
- Earth and apace exploration

Enterprise application
- Email, documents, files
- Application log
- Transaction Records

Web 2.0 Data
- Search log/click stream
- Twitter/Blog/SNS
- Wiki

Other unstructured data
- Video/Movie
- Graphics
- Digital widgets

Bigger Challenges

Scale out automatically
- Vs. Scale up manually

More capacity, bigger pool
- E.g., 10 PB in a single file system

New process capability
- Loading, Analyzing, Moving data
- Intelligence

Better performance
- Linear vs. exponent
  Faster

Autonomous
- Fewer human interface
- Lower cost

Data Management in the cloud

- Two components of data management market
  - Transactional Data Management (OLTP)
    - Banks, airline reservation, online e-commerce
    - ACID, write-intensive
  - Analytical Data Management (OLAP)
    - Business Intelligence, decision support
    - Query-intensive
- Challenges of data management in the cloud
  - Scalability
  - Fault Tolerance
  - Availability & Consistence
  - Transaction Management
  - Flexible Schema

Cloud Database

- Data analytics in the cloud
  - Parallel DBMS
  - MapReduce
- Transactional data management in the cloud
  - NoSQL Store
  - SQL Database
- Cloud data service (Data as a service)
  - Multi-tenant data management
  - Auto-administration

What Is Hadoop?

"Flexible and available architecture for large scale computation and data processing on a network of commodity hardware"

"Hadoop is like a Parallel DBMS.

But Hadoop is not Parallel DBMS"

Parallel DBMS technologies

- Share-nothing nodes cluster
- Relational Data Model
- Indexing
- Familiar SQL interface
- Parallel query execution
  - Horizontal partitioning of relational tables with partitioned execution of SQL queries
- Advanced query optimization

MapReduce vs Parallel DBMS

	Parallel DBMS	MapReduce
Schema Support	√	Not out of box
Indexing	√	Not out of box
Programming Model	Declarative (SQL)	Imperative (C/C++, Java, …)
Optimizations (Compression, Query Optimization)	√	Not out of box
Flexibility	Not out of box	√
Fault Tolerance	Coarse grained techniques	√

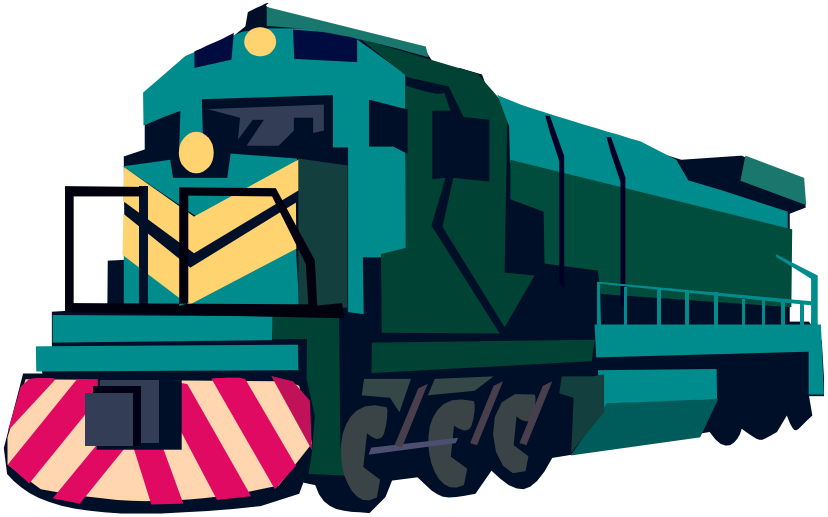
Use The Right Tool For The Right Job

Relational Databases:

When to use?

- **Interactive Reporting (<1sec)**
- **Multistep Transactions**
- **Lots of Inserts/Updates/Deletes**

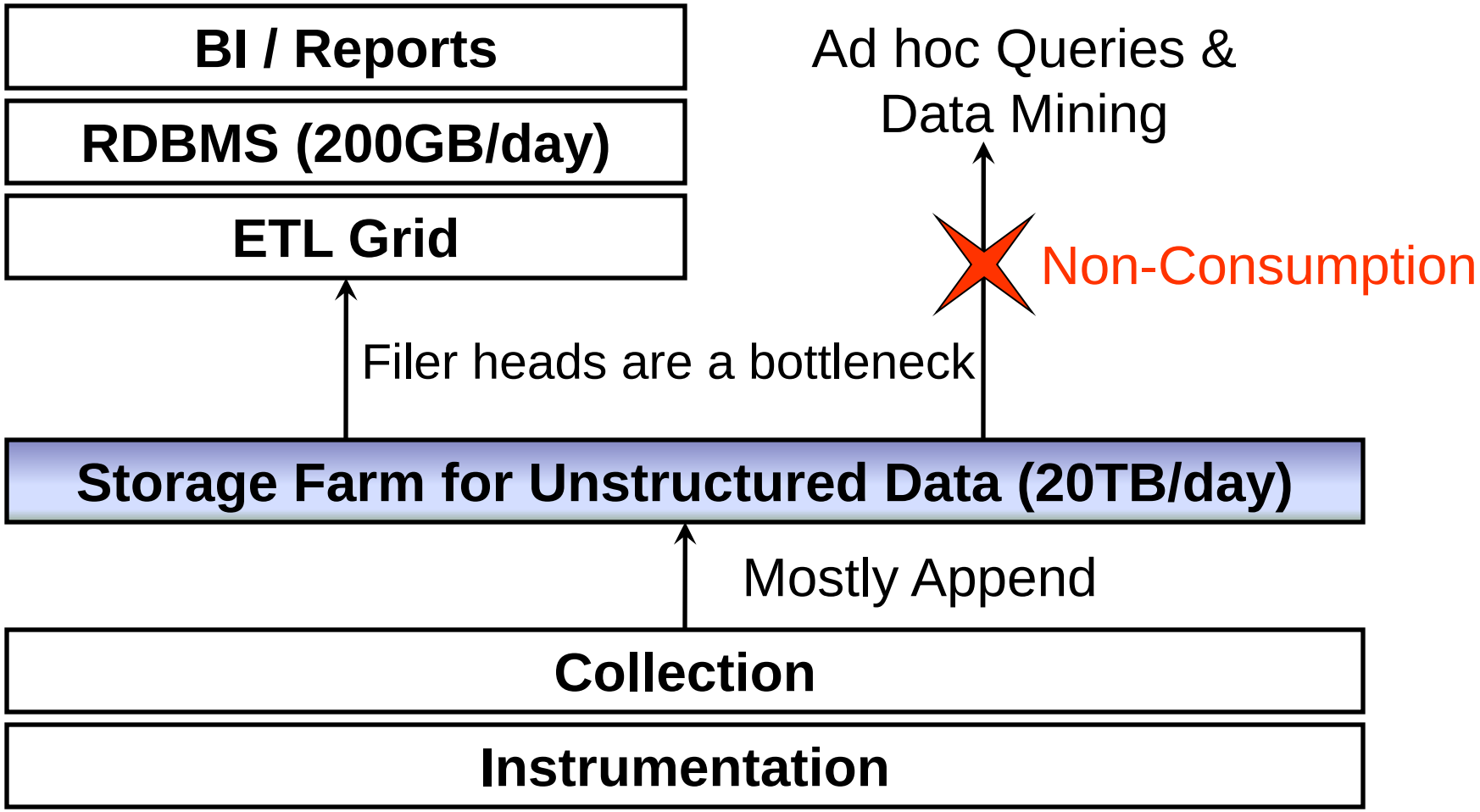
Hadoop:



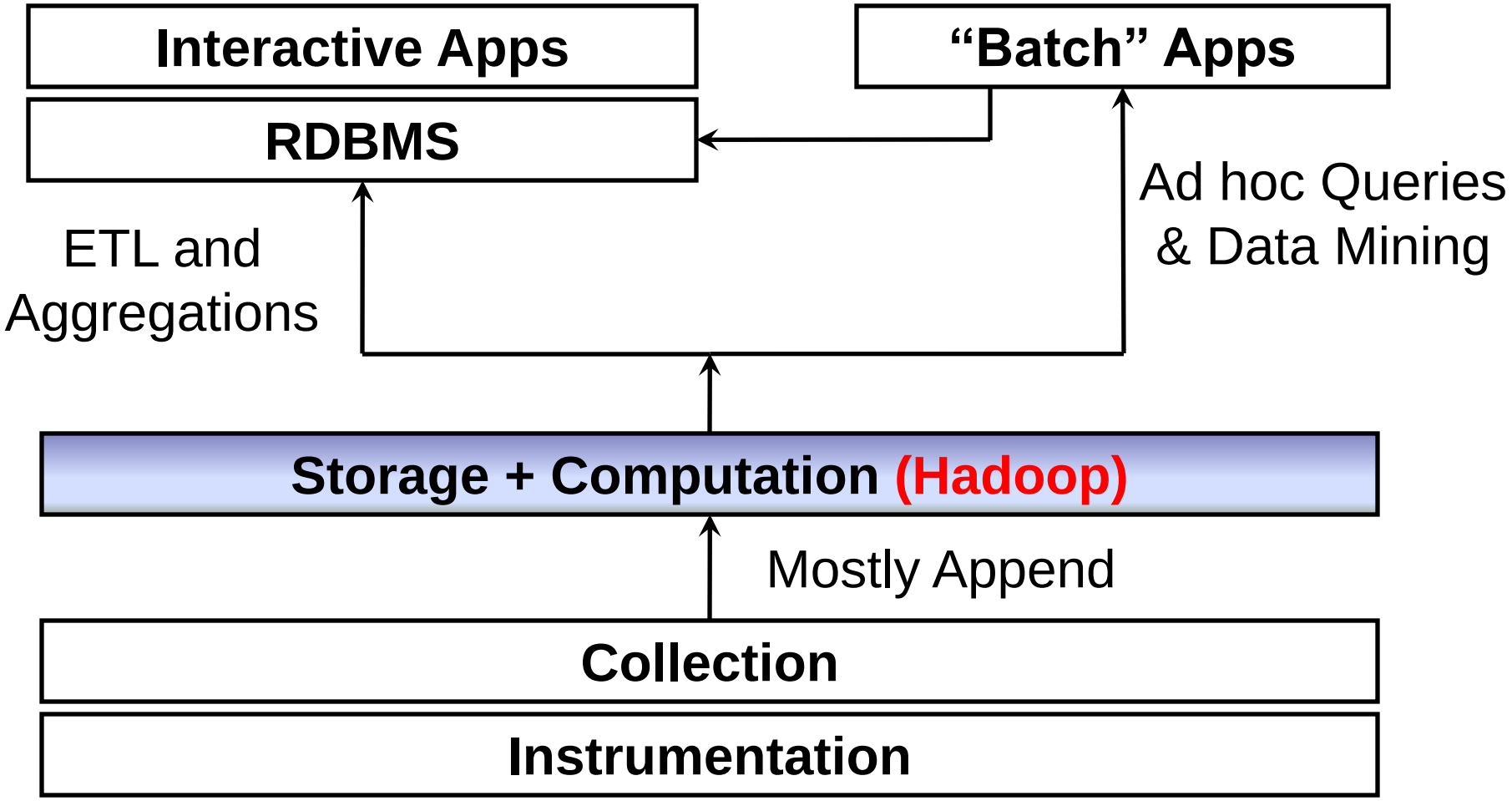
When to use?

- **Affordable Storage/Compute**
- **Structured or Not (Agility)**
- **Resilient Auto Scalability**

Older BI Systems for Limited Raw Data Access



The New Solution: A Store-Compute Grid



Hadoop Is More Than Just Analytics/BI

- Searching
- Log processing
- Recommendation systems
- Fraud Detection and Fighting Email Spam
- Collaborative Filtering
- Video and Image analysis
- Gene Sequence Alignment

It seems that Hadoop can do anything.

But it doesn't

Hadoop Criticisms (part 1)

- **Hadoop can't do quick random lookups**
  - HBase enables low-latency key-value pair lookups (no fast joins)
- **Hadoop doesn't support updates/inserts/deletes**
  - Not for multi-row transactions, but HBase enables transactions with row-level consistency semantics
- **Hadoop isn't highly available**
  - Though Hadoop rarely loses data, it can suffer from down-time if the master NameNode goes down. This issue is currently being addressed, and there are HW/OS/VM solutions for it

Hadoop Criticisms (part 2)

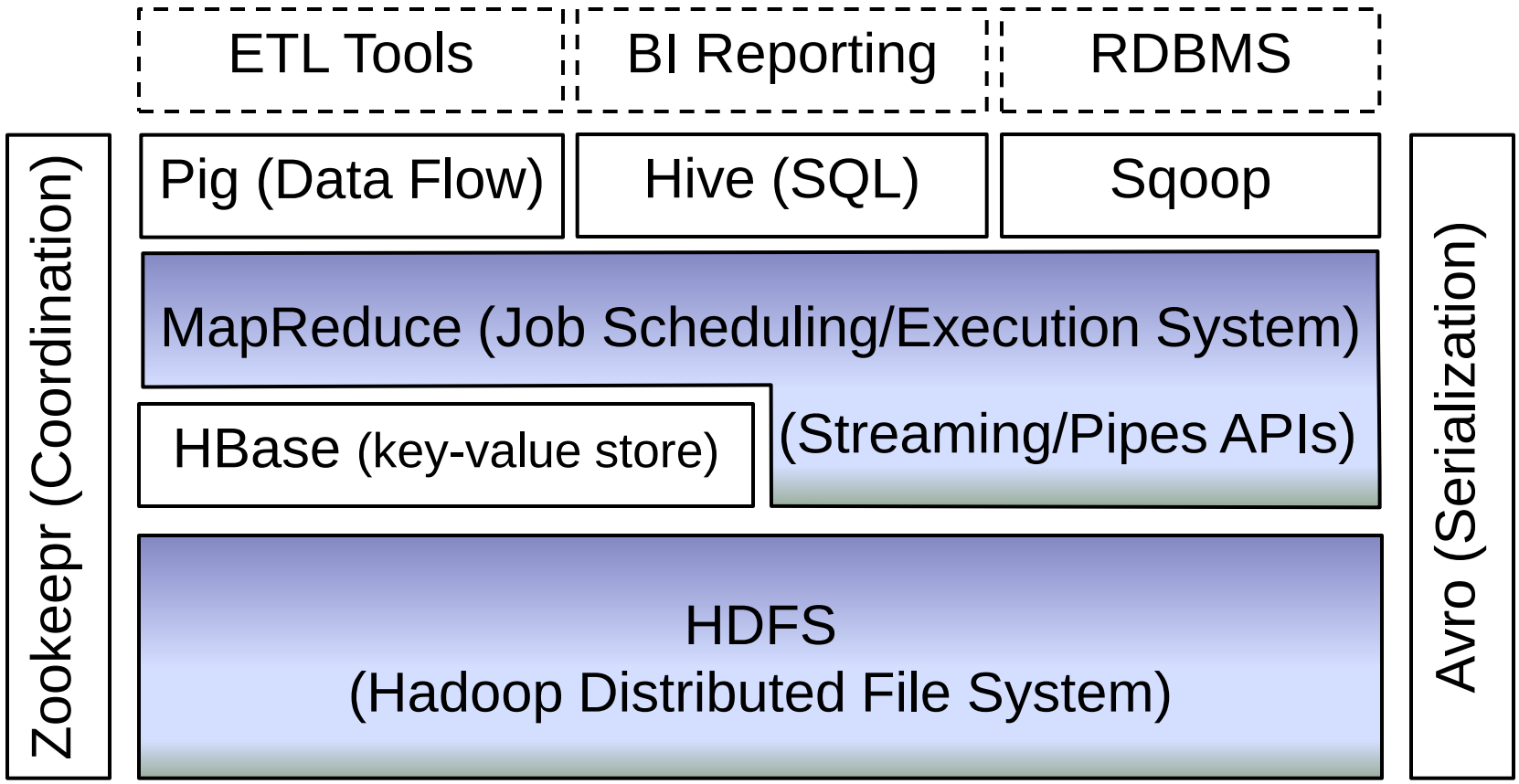
- **Hadoop can't be backed-up/recovered quickly**
  - HDFS, like other file systems, can copy files very quickly. It also has utilities to copy data between HDFS clusters
- **Hadoop can't support data flow**
- **Hadoop can't talk to other systems**
  - Hadoop can talk to BI tools using JDBC, to RDBMSes using Sqoop, and to other systems using FUSE, WebDAV & FTP

"There is a trend for Combining both Hadoop and Parallel DBMS"

Typical Research on hybrid solution

- HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads
- FlexDB: A cloud-scale database engine based on Hadoop

Apache Hadoop Ecosystem



Hive

- Data Warehouse infrastructure that provides data summarization and ad hoc querying on top of Hadoop
- MetaStore
- Hive Query Language
  - Basic SQL: Select, From, Join, Group By
  - Equi-Join, Multi-Table Insert, Multi-Group-By
  - Batch query

Pig

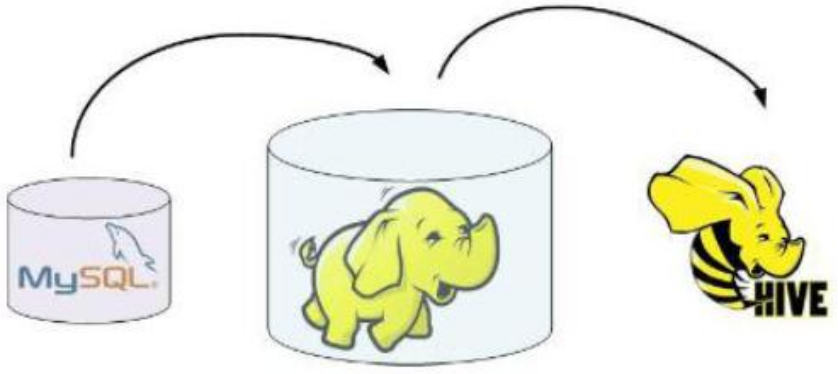
- A high-level data-flow language and execution framework for parallel computation
- Simple to write MapReduce program
- Abstracts you from specific detail
- Focus on data processing
- Data flow
- For data manipulation

PIG Language Example

```
Users = LOAD ‘users’ AS (name, age);
Fltrd = FILTER Users BY
          age >= 18 AND  age <=25;
Jnd = JOIN Fltrd BY name, Pages BY user;
Grpd = GROUP Jnd BY url;
…….
```

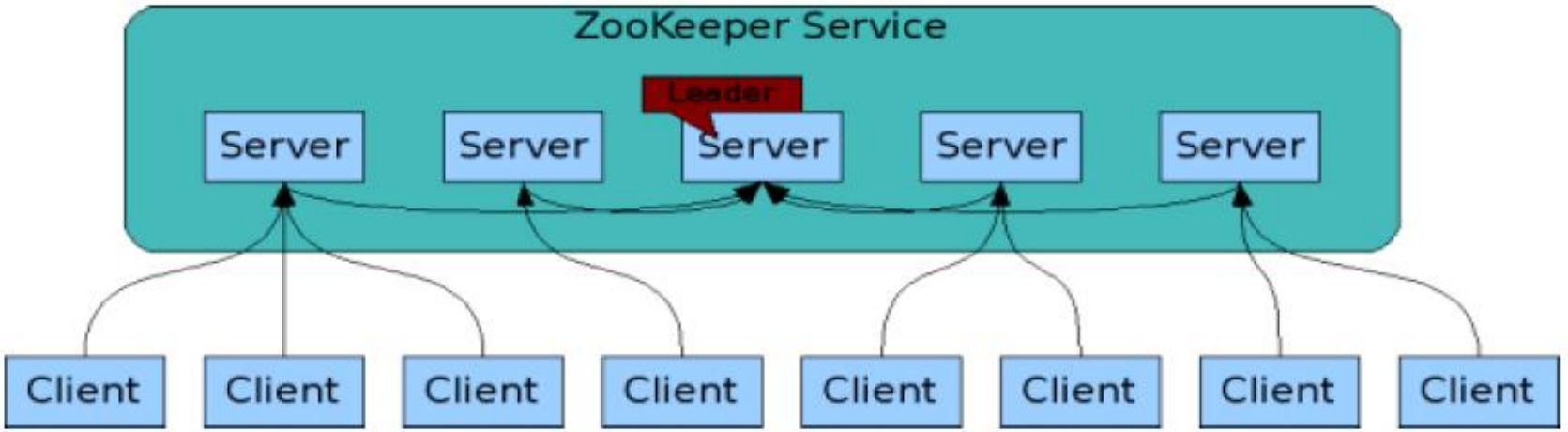
Sqoop

- Sqoop is a tool designed to help users of large data import existing relational databases into their Hadoop clusters
- Automatic data import
- SQL to Hadoop
- Easy import data from many databases to Hadoop
- Generates code for use in MapReduce applications
- Integrates with Hive



Zookeeper

- A high-performance coordination service for distributed applications
- Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

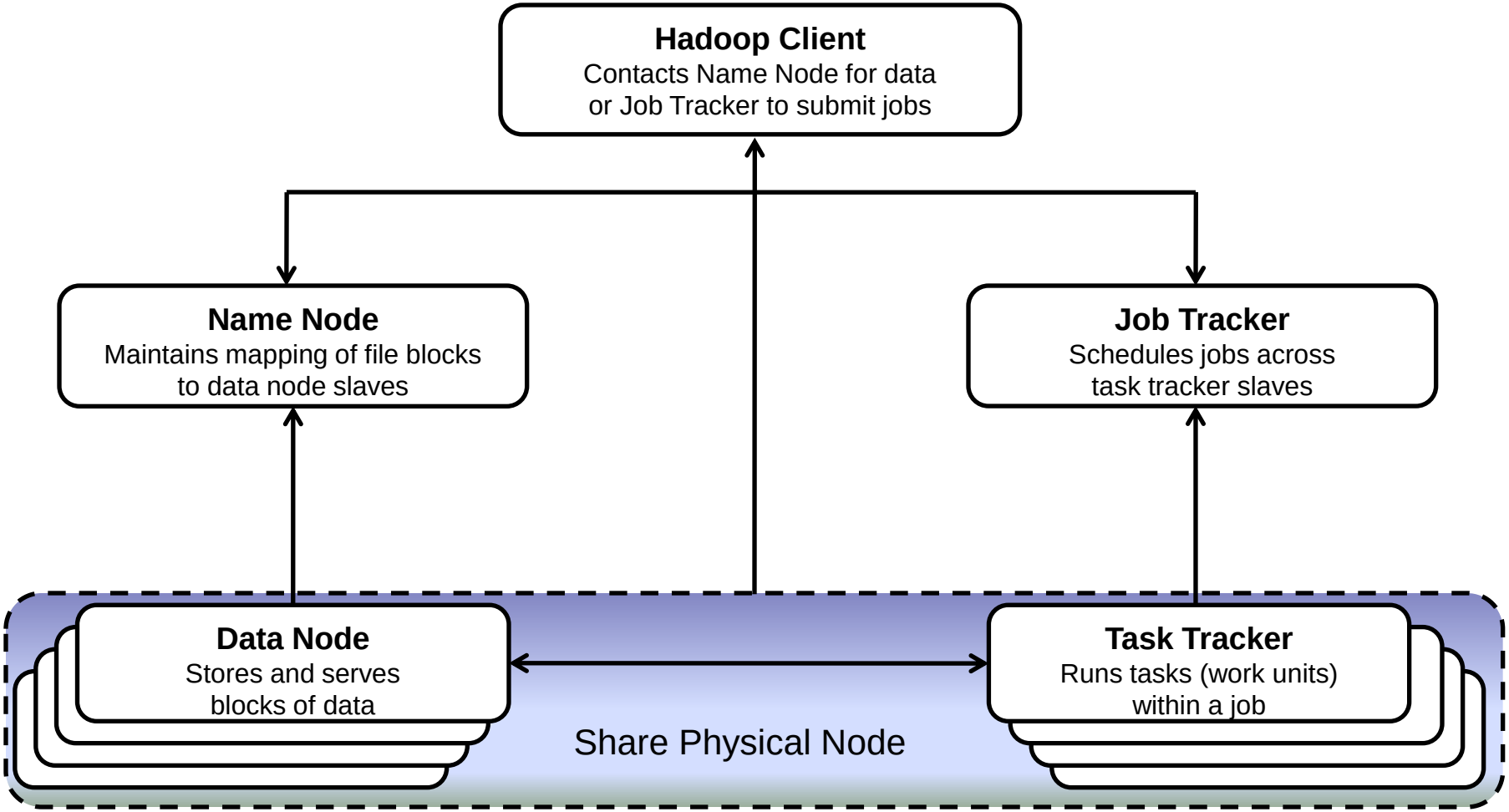


- A leader is elected at startup
- Follower service clients, all updates go through leader
- Update responses are sent when a majority of servers have persisted the change

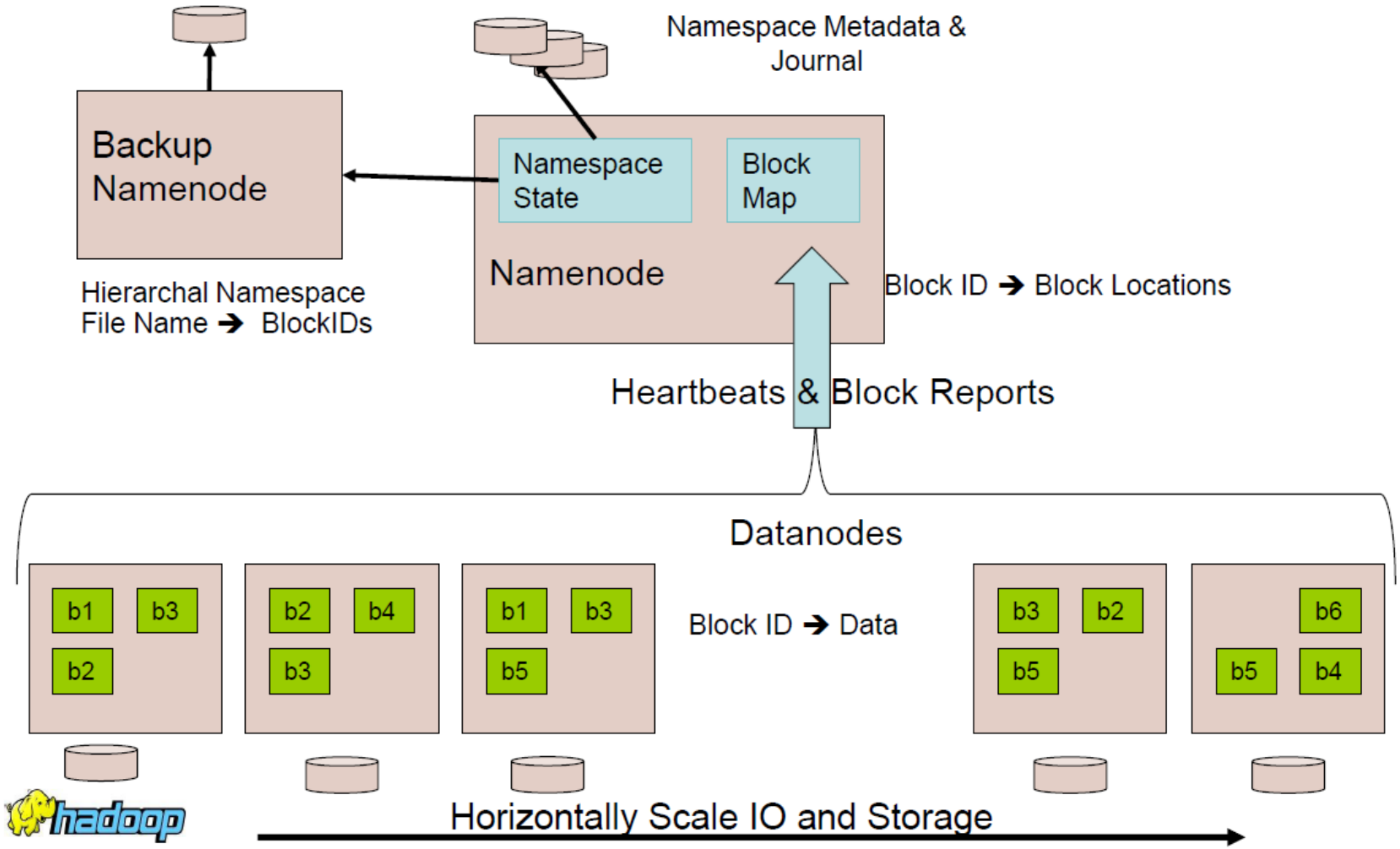
Future

- HDFS Federation
- Next generation MapReduce
- High Availability

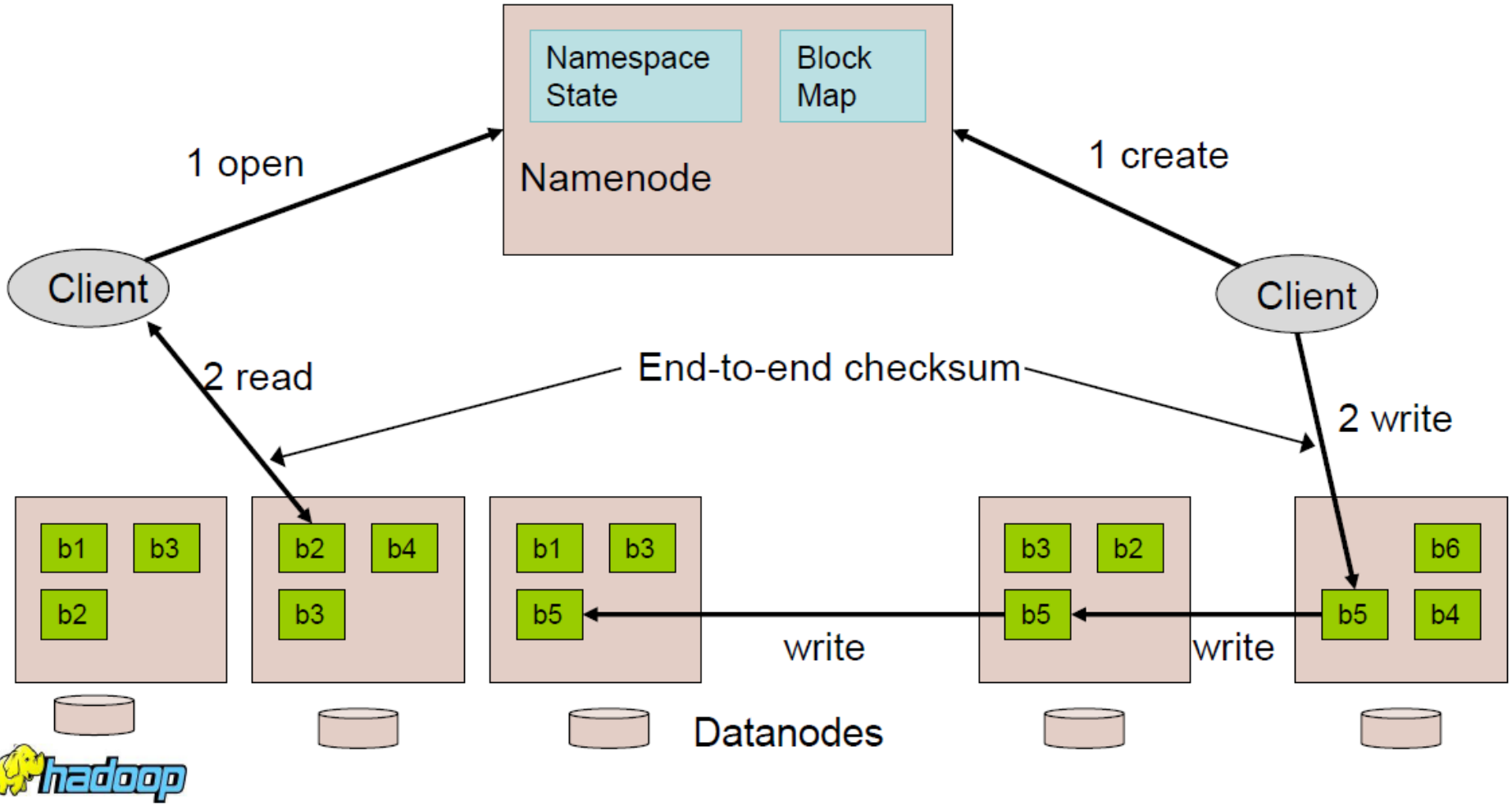
Hadoop1.0


Hadoop Client
Contacts Name Node for data
or Job Tracker to submit jobs
Name Node
Maintains mapping of file blocks
to data node slaves
Job Tracker
Schedules jobs across
task tracker slaves
Data Node
Stores and serves
blocks of data
Task Tracker
Runs tasks (work units)
within a job
Share Physical Node

HDFS


Namespace Metadata &
Journal
Backup
Namenode
Namespace
State
Block
Map
Namenode
Hierarchal Namespace
File Name ➔ BlockIDs
Block ID ➔ Block Locations
Heartbeats & Block Reports
Datanodes
b1
b3
b2
b2
b4
b3
b1
b3
b5
Block ID ➔ Data
b3
b2
b5
b6
b5
b4
hadoop
Horizontally Scale IO and Storage

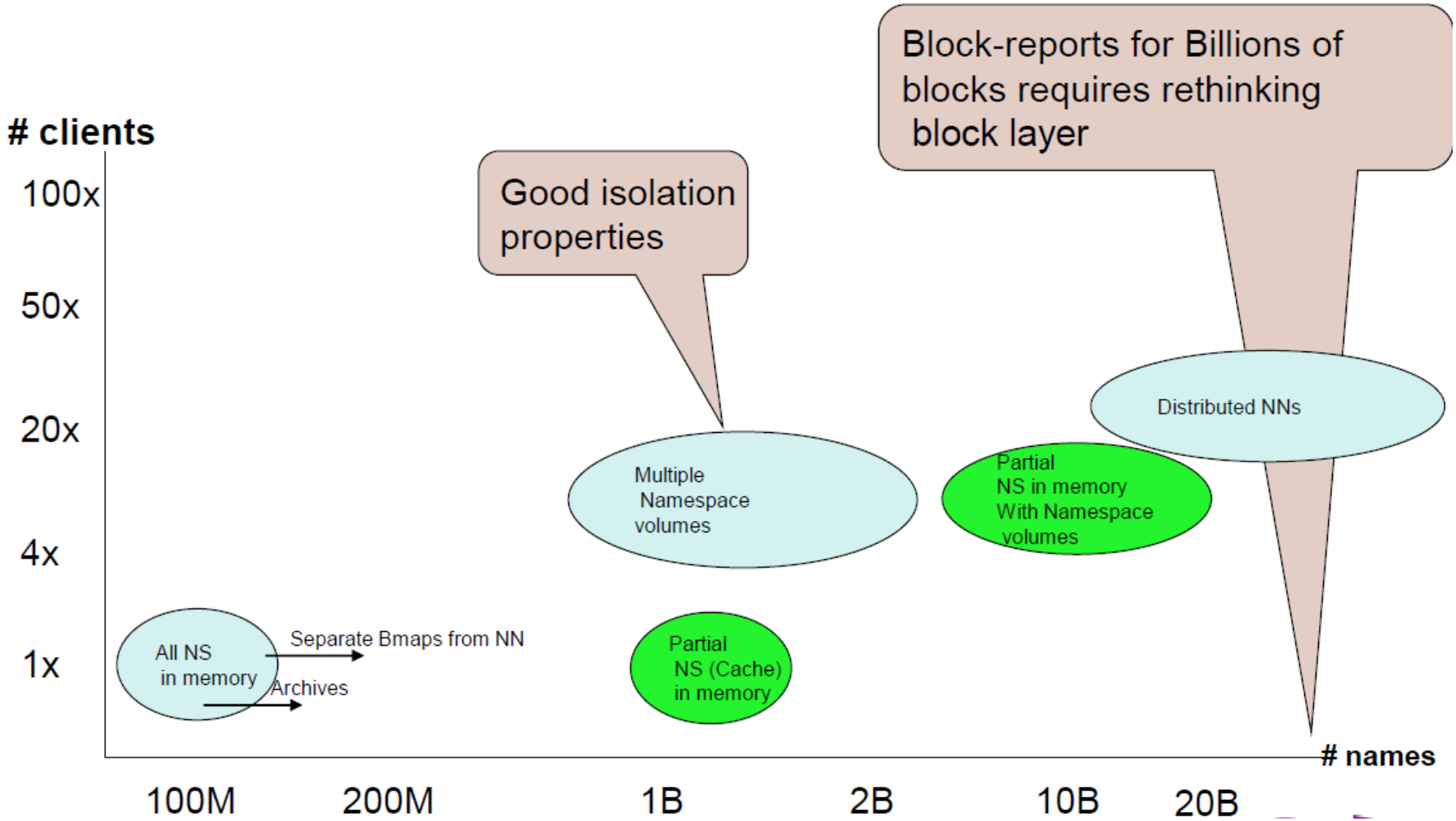
HDFS Client reads and writes



Current Limitations of Hadoop NameNode

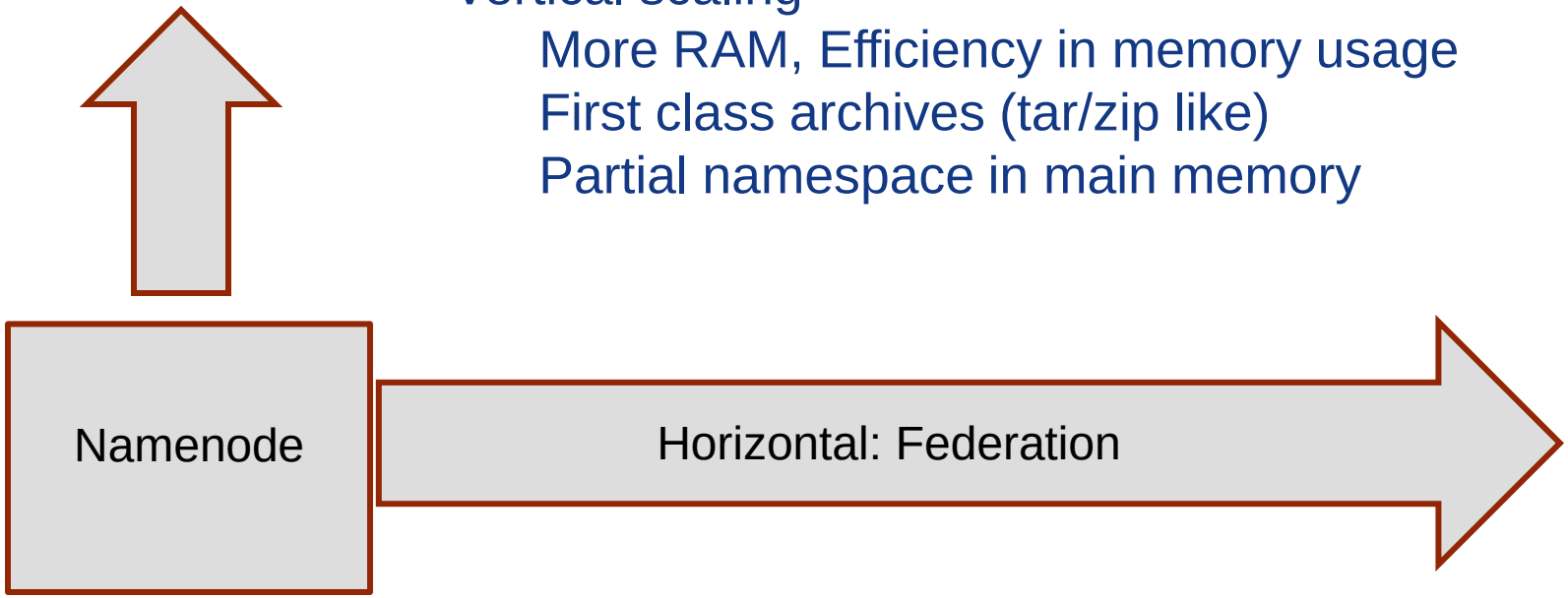
- Early Gains
  - Simple design allowed rapid improvements
    - Namespace is all in RAM, simper locking
    - Improved memoery usage in 0.16, JVM Heap configuration
- Growth of number of files and storage is limited by adding RAM to namenode
  - 50G heap = 200M “fs objects” = 100M names + 100M Blocks
    - 14PB of storage (50MB blocksize)
    - 4000 nodes
- Goal:
  - Clusters of 6000 nodes, 100,000 cores & 10K concurrent jobs, 100PB raw storage per cluster

Scaling the Name Service: Options



Vertical vs Horizontal

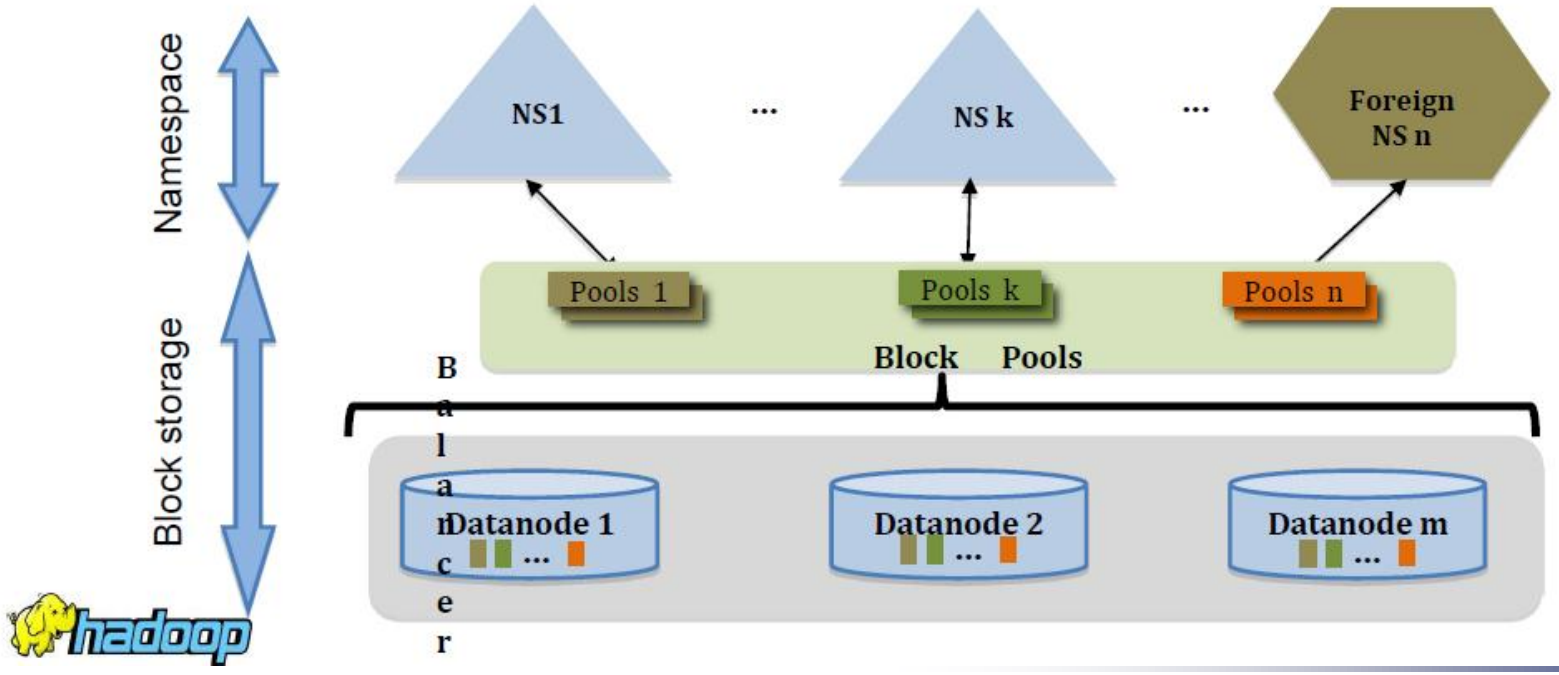
Vertical scaling
- More RAM, Efficiency in memory usage
- First class archives (tar/zip like)
- Partial namespace in main memory



Horizontal scaling/federation benefis:
- Scale
- Isolation, Stability, Availability
- Flexibility
- Other Namenode implementations or non-HDFS namespaces

Block Storage Subsystem

- Shared storage provided as pools of blocks
- Namespaces (HDFS, others) use one or more block-pools
- Note: HDFS has 2 layers today – we are generalizing and extending it

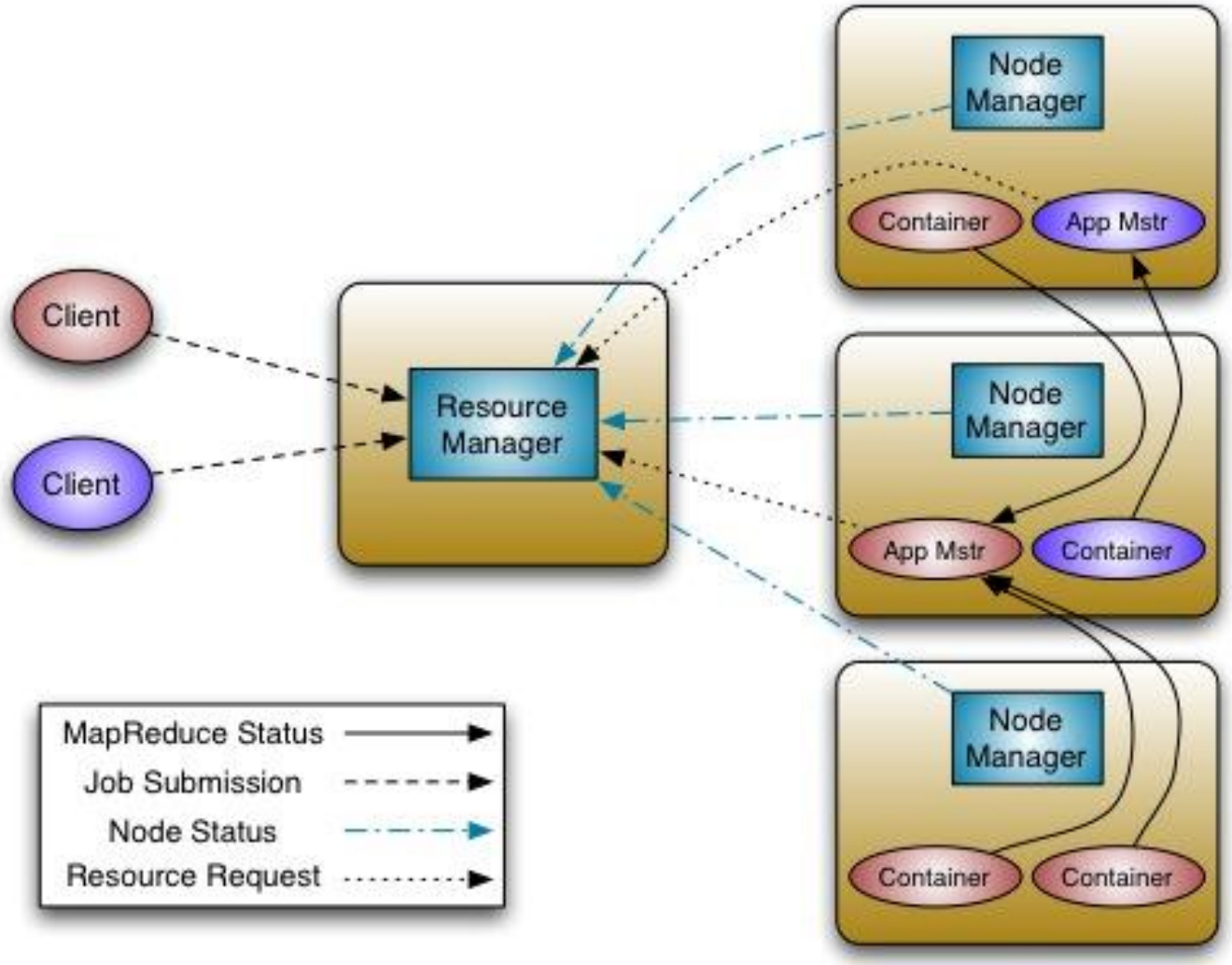


Next Gen MapReduce

Current Limitations of Hadoop JobTracker

- Scalability
  - Maximum cluster size – 4000 nodes
  - Maximum concurrent tasks – 40000
  - Coarse synchronization in JobTracker
- Single point failure
- Restart is very tricky due to complex state
- Lack support for alternate paradigms
- Lack of wire-compatible protocols

Next Generation MapReduce



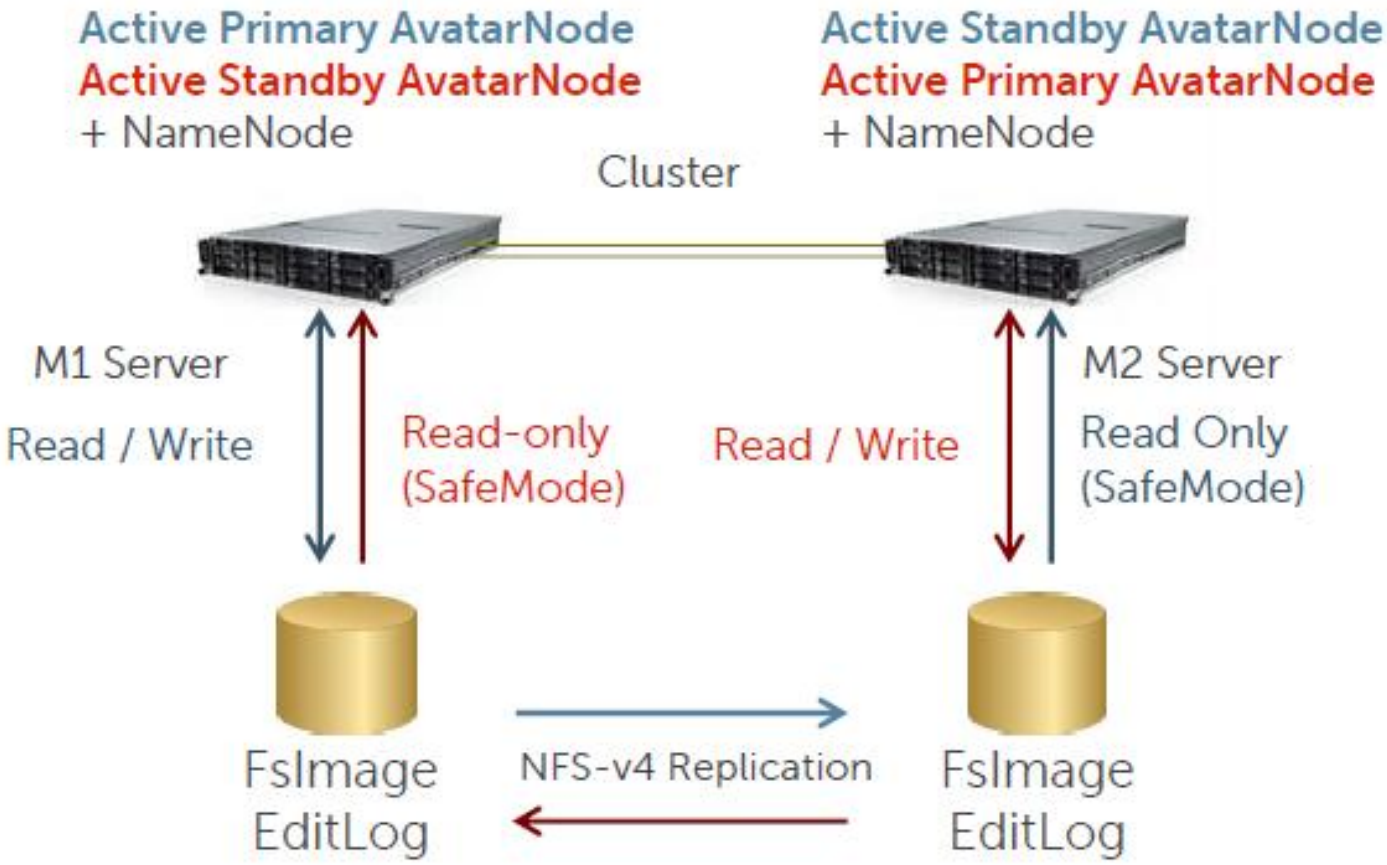
Architecture

- Resource Manager
  - Global resource scheduler
  - Hierarchical queues
- Node Manager
  - Per-machine agent
  - Manages the life-cycle of container
  - Container resource monitoring
- Application Master
  - Per-application
  - Manages application scheduling and task execution

High Availability Node: Avatar Node

- HDFS client are configured to access the AvatarNode via a Virtual IP Address(VIP)
- When Primary AvatarNode is down, the Standby AvatarNode takes the relay
- The Standby AvatarNode ingests all committed transactions because it reopens the edits log and consumes all transactions until the end of the file
- The Standby AvatarNode finished ingestion of all transactions from the shared NFS filer and then leaves SafeMode
- The VIP switches from Primary AvatarNode to Standby AvatarNode

Avatar Node


Active Primary AvatarNode
Active Standby AvatarNode
+ NameNode
Active Standby AvatarNode
Active Primary AvatarNode
+ NameNode
Cluster
M1 Server
Read / Write
Read-only
(SafeMode)
Read / Write
M2 Server
Read Only
(SafeMode)
FsImage
EditLog
NFS-v4 Replication
FsImage
EditLog

Who Research Hadoop

Berkeley
UNIVERSITY OF CALIFORNIA

Washington
University in St.Louis

Massachusetts
Institute of
Technology

Yale
UNIVERSITY

Who Used Hadoop

facebook

FOX

YAHOO!

Google

twitter

YouTube

IBM

淘宝网 Taobao.com
阿里巴巴旗下网站

Research Scopes and Topics in Big Data

- Search and Analytics
  - Search: Entity Search, Faceted Search, Associative Search
  - Analytics: Text Analysis, Activity Modeling and Sequence Analysis, Real-time Data Analysis for Streaming, Parallel Data Mining Algorithms
- MPP Databases and Data Services
  - Parallel Database: Parallel Query Optimization, Data Partitioning and Replication, Distributed Transaction
  - In-memory Database: Cache, Recovery, Consistence
  - Database as a Service: Multi-tenant Data Management, Auto-Administration
- Hadoop/NoSQL
  - Hadoop: Single-node Failure, Performance, Real-time MapReduce Scheduler and Fault Tolerance
  - NoSQL: Key-Value Store, Documents Store, Graph Data Store

Takeaways

- What is Hadoop?
  - Hadoop is like a DBMS, but not DBMS
- Hadoop is powerful, but not powerful enough
- The future of the Hadoop

Big Data is a Big Deal.

The challenges are clear but the opportunities are abundant.