Business Intelligence for Big Data

Will Gorman, Vice President, Engineering

May, 2011

What is BI?

Business Intelligence =

reports, dashboards, analysis, visualization, alerts, auditing, data transformation

Hadoop and BI

Example Hadoop BI Use Cases Today

Transactional

- Fraud detection
- Financial services/stock markets

Sub-Transactional

- Weblogs
- Social/online media
- Telecoms events

Example Hadoop BI Use Cases Today

Non-Transactional

- Web pages, blogs etc
- Documents
- Physical events
- Application events
- Machine events

In most cases structured or semi-structured

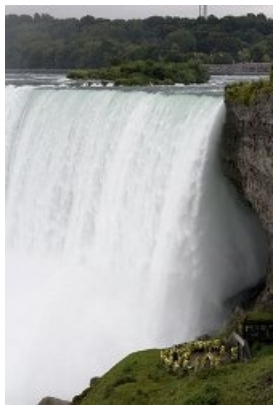
Traditional BI

Data Mart(s)

Data Source

? ? ? ? ? ? ?

Tape/Trash



Data Lake

- Single source
- Large volume
- Not distilled
- Can be treated

Data Lakes

- 0-2 data lakes per company
- Known and unknown questions
- $1-10k questions, not $1m ones
- Multiple user communities
- Don’t fit in traditional RDBMS with a reasonable cost

Data Lake Requirements

- Store all the data
- Satisfy routine reporting and analysis
- Satisfy ad-hoc query / analysis / reporting
- Balance performance and cost

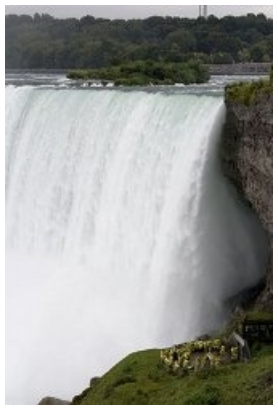
Traditional BI

Data Mart(s)

Data Source

? ? ? ? ? ? ?

Tape/Trash



Big Data Architecture

Data Mart(s)

Ad-Hoc

Data Warehouse

Data Lake(s)

Data Source

Does Hadoop Replace Data Marts?

- If it behaves like database
- If it has low latency (sub-second)

Hadoop (to date)

- Databases (Hive) are immature
- Some databases are no-SQL

BI Tools Need...

Structured Query Language

Why Hadoop and BI?

- Distributed, scalable file system
- Distributed processing
- Commodity hardware
- Scales out beyond technology and/or economy of a RDBMS

In many cases it's the only viable solution

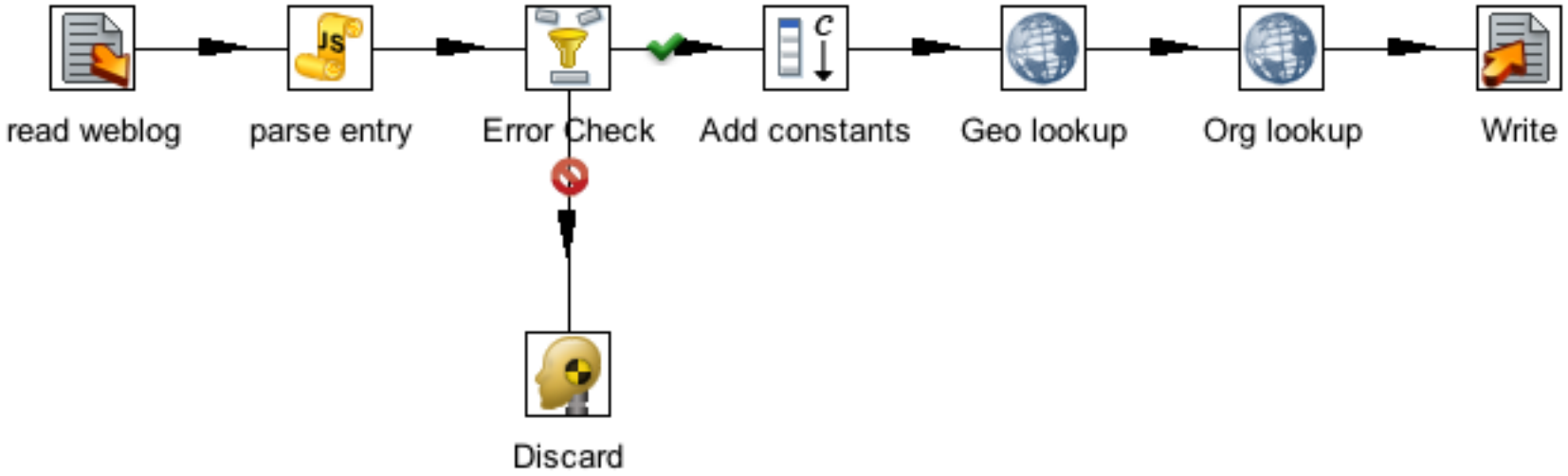
Hadoop and BI?

"The working conditions within Hadoop are shocking"

ETL Developer

Hadoop and BI?

Instead of this...



Hadoop and BI?

You have to do this...

```
public void map(
    Text key,
    Text value,
    OutputCollector output,
    Reporter reporter)

public void reduce(
    Text key,
    Iterator values,
    OutputCollector output,
    Reporter reporter)
```

MapReduce Limitations

Doing everything with MapReduce is like doing everything with recursion.

You can do it, but that doesn't mean it's going to be easy

Google's Use Case

- Needed to index the internet
- Huge set of unstructured data
- Predetermined input
- Predetermined output (the index)
- Predetermined questions
- Single user community
- Needed parallel processing and storage

Their answer was MapReduce (MR)

Yahoo's Use Case

- Needed to index the internet
- Huge set of unstructured data
- Predetermined input
- Predetermined output (the index)
- Predetermined questions
- Single user community
- Needed parallel processing and storage

Their answer was Hadoop (w/ MapReduce)

Current Use Cases

- ✗ Not indexing the internet
- ✗ Huge set of semi/structured data
- ✗ Different input source and format
- ✗ Different outputs
- ✗ Different questions
- ✗ Multiple user communities
- ✓ Need parallel processing and storage

Unfortunately Hadoop
wasn't designed
for most BI requirements

Hadoop's Strengths and Weaknesses

- Distributed processing
- Distributed file system
- Commodity hardware
- Scales out beyond technology and/or economy of a RDBMS

But...

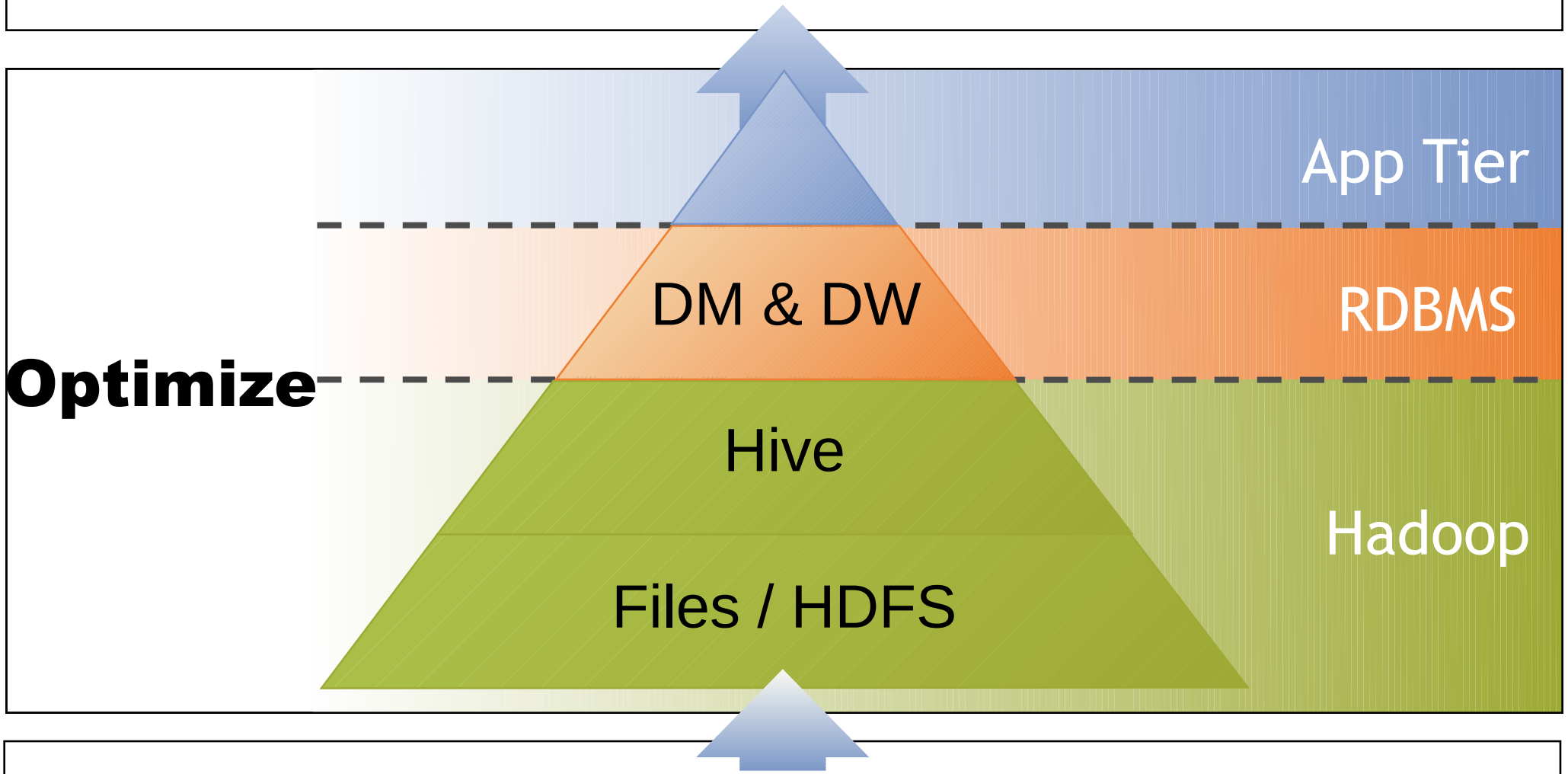
- Not designed for BI

BI and Hadoop Architecture

Until Hadoop behaves and performs like a database a hybrid architecture is needed

- Data sources
- Hadoop
- Data marts
- BI tools

Visualize Reporting / Dashboards / Analysis



Load Applications & Systems

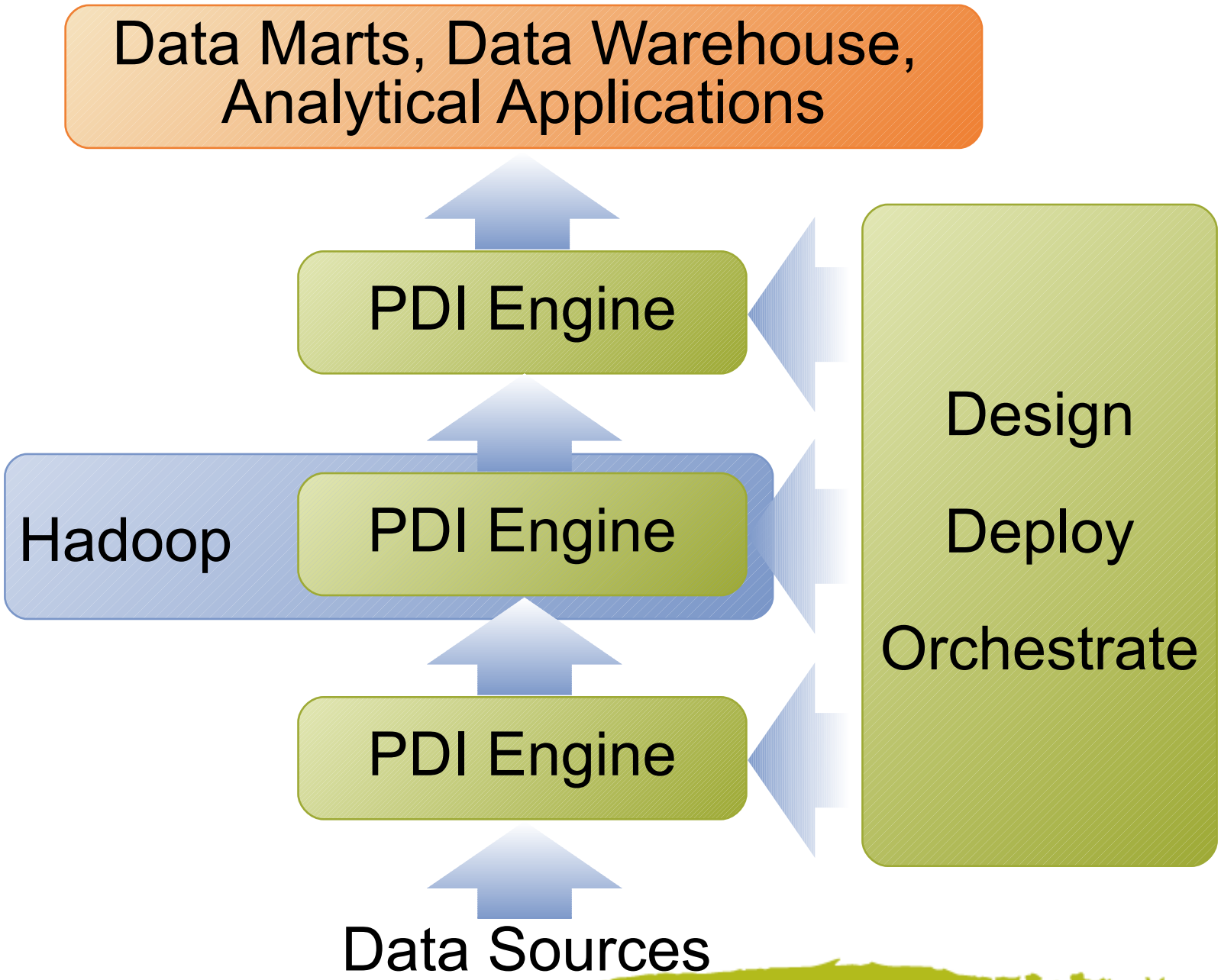
Why not add to Hadoop the things it's missing...

... until it can do

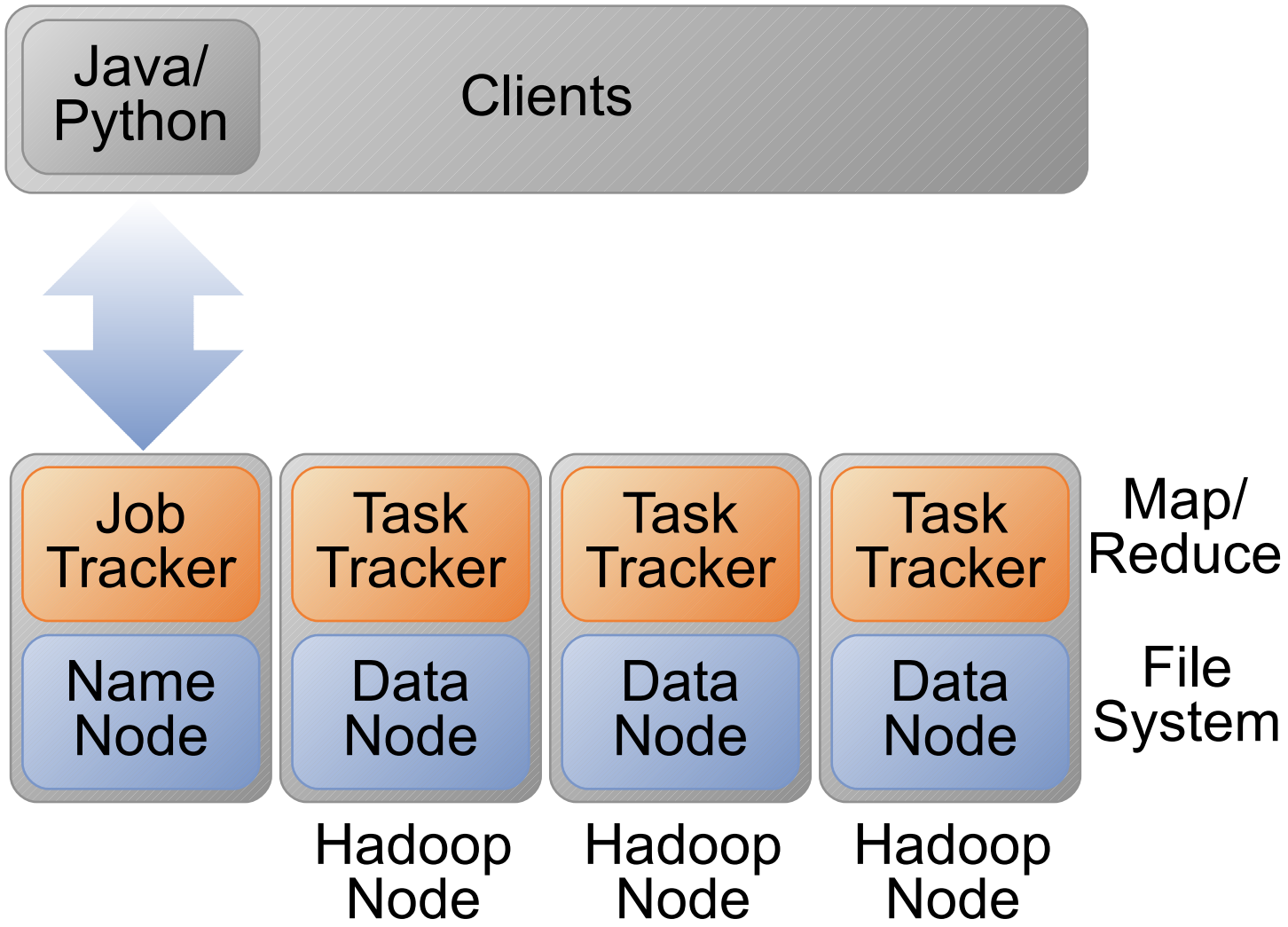
what we need it to?

If only we had a
Java, embeddable,
data transformation engine...

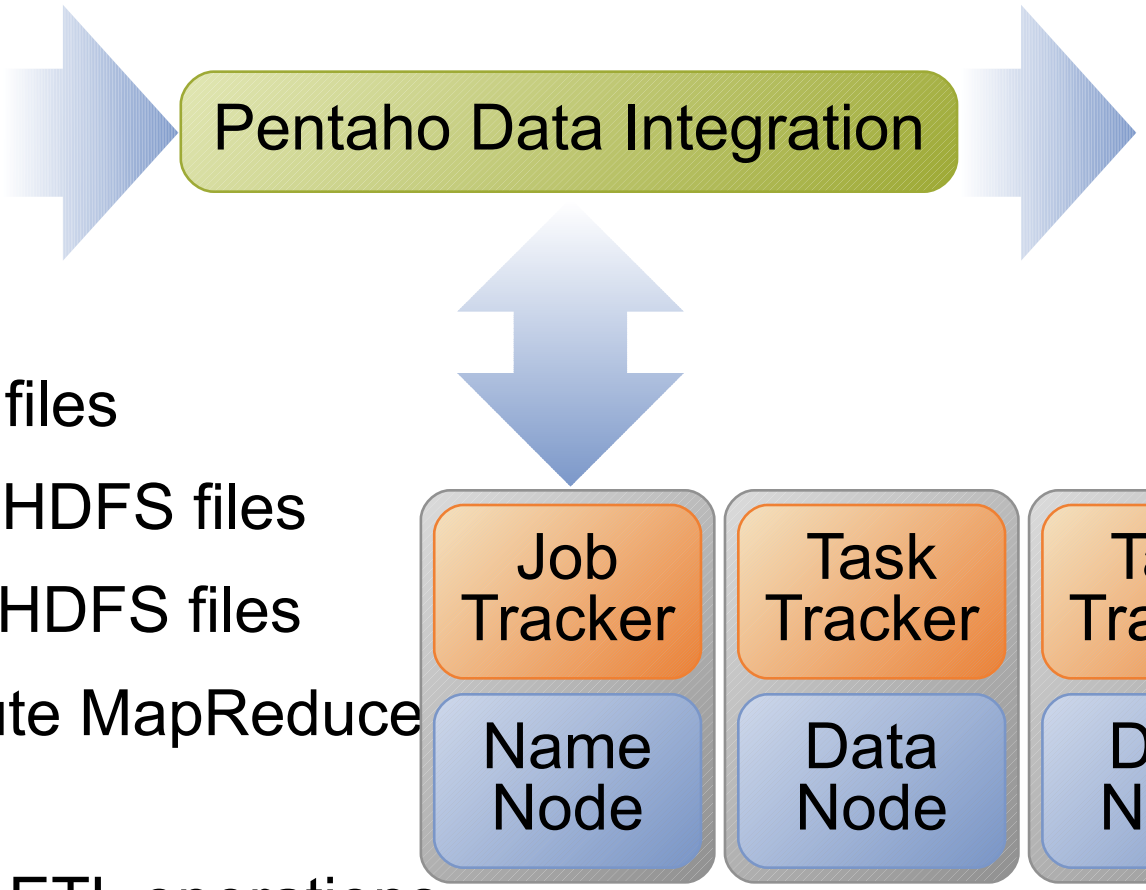
Pentaho Data Integration



Hadoop Architecture



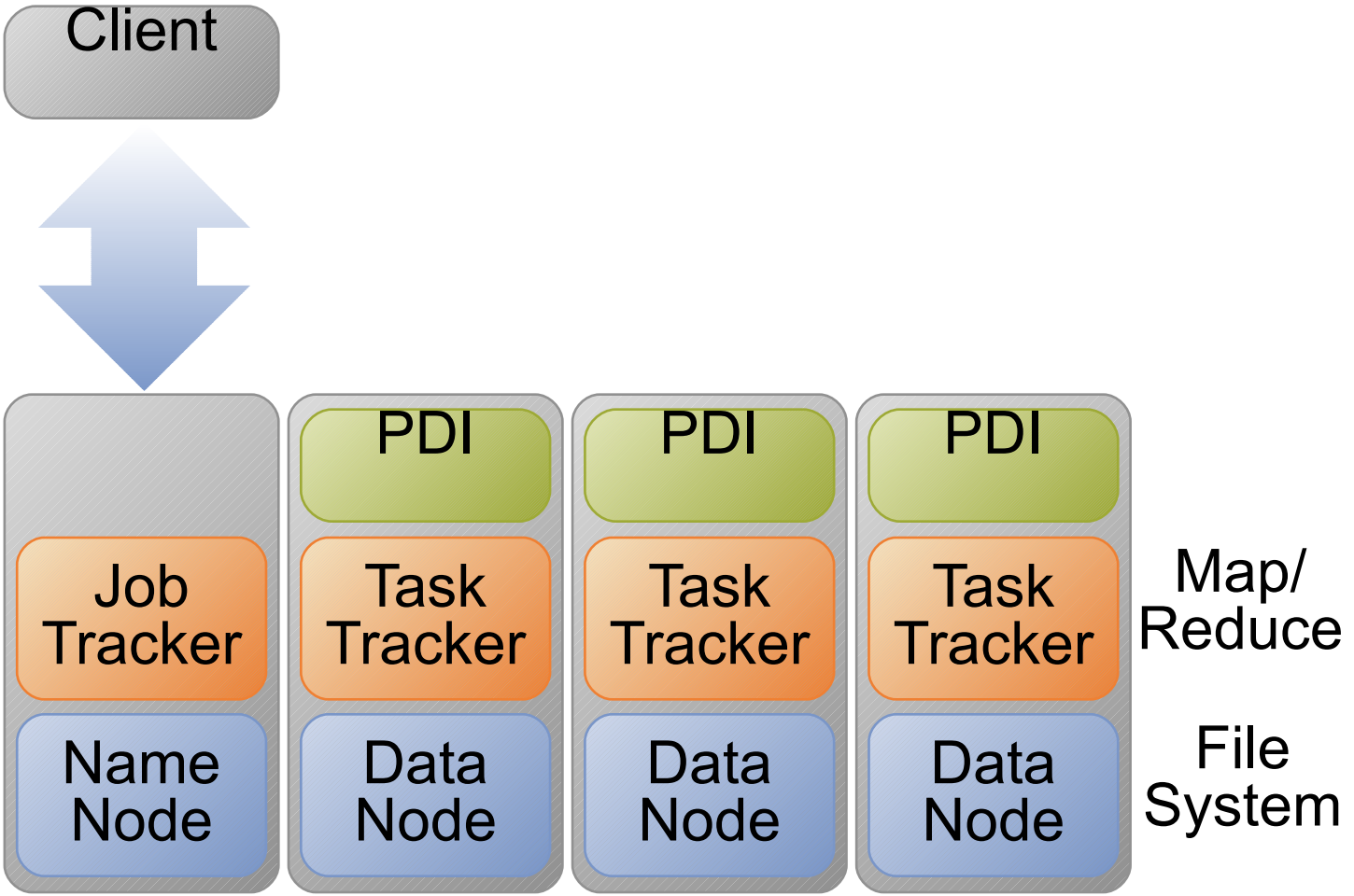
Pentaho/Hadoop Architecture - External



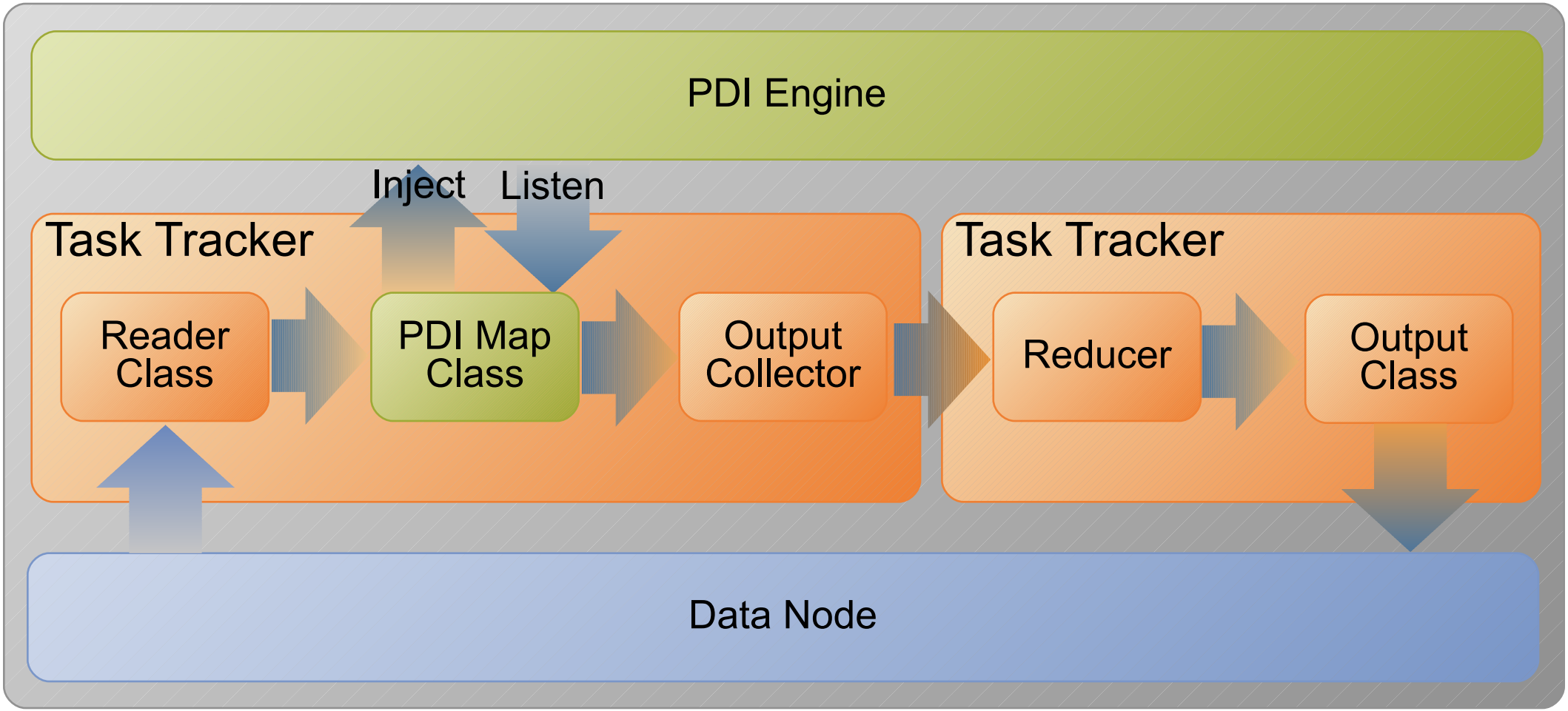
- Move files
- Read HDFS files
- Write HDFS files
- Execute MapReduce jobs
- Other ETL operations

Pentaho/Hadoop Architecture - Internal

- Exec ETL in parallel

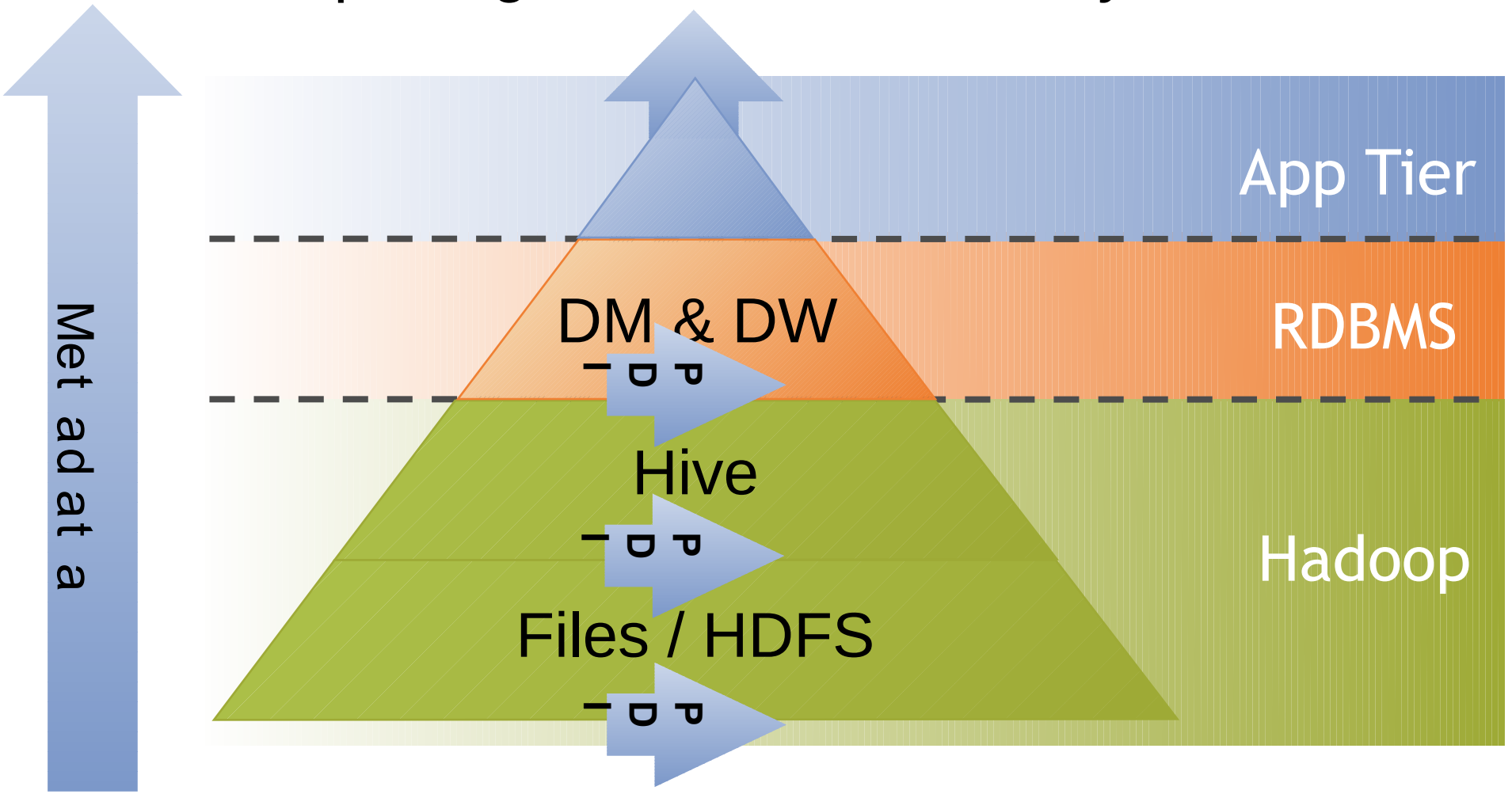


Pentaho/Hadoop Architecture - Internal



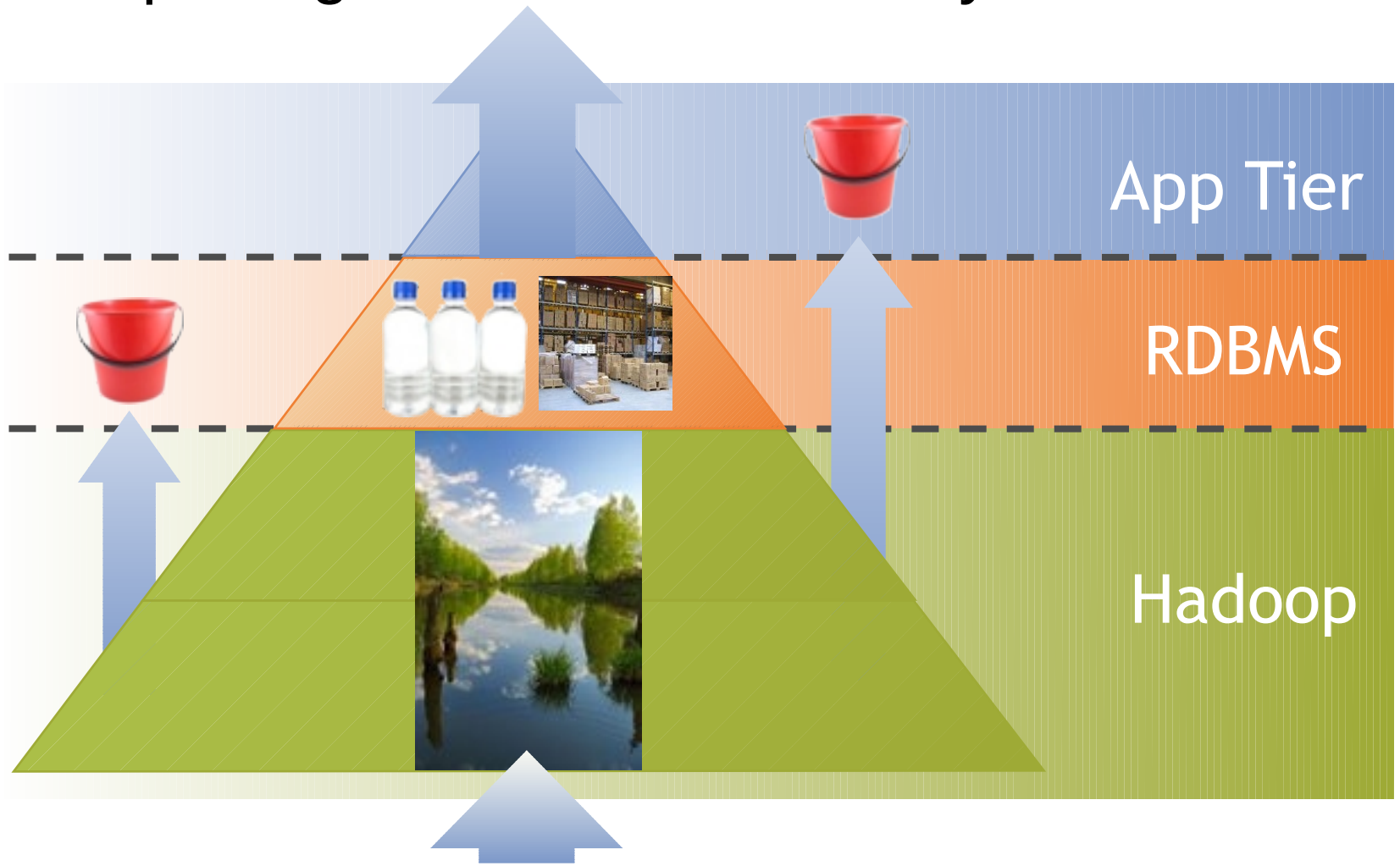
The PDI Engine executes within the Task Tracker JVM
The PDI Engine can also execute as a Reduce task

Reporting / Dashboards / Analysis



Applications & Systems

Reporting / Dashboards / Analysis



Applications & Systems

Demo

FAQ

1. Will Pentaho contribute to Apache's Hadoop projects? ***Yes***
2. Will Pentaho distribute Hadoop as part of their product? ***Unlikely***
3. What version of Hadoop will be supported? ***Initially 20.2***
4. Will Pentaho's APIs allow existing open source APIs to be used in parallel? ***Yes***

FAQ

5. Will Pentaho provide support or services to help setup Hadoop? ***Yes, no, maybe***

6. What are the requirements to be in the Pentaho Hadoop beta program?

Requirements, be serious, have started already, etc

Side Topic:

No-SQL and BI

BI Tools Need...

Structured Query Language

For Modeling...

- Data rich
- Metadata poor
- Sample = table scan
- Pre-emptive attribute selection

BI Tools Don’t Need

- CREATE / INSERT
- UPDATE
- DELETE
- (only Read needed)
- No ACID transactions

Mondrian (OLAP) Needs

Required:

- SELECT
- FROM
- WHERE
- GROUP BY
- ORDER BY

Nice to have:

- HAVING
- ORDER BY ... NULLS COLLATE
- COUNT(DISTINCT x,y)
- COUNT(DISTINCT x), COUNT(DISTINCT y)
- VALUES (1,'a'), (2,'b')

Side Topic:

Hadoop and Data Warehouses

Can I Use Hadoop as a Data Warehouse?

Yes, probably

Should I Use Hadoop as a Data Warehouse?

No, probably not*

* until performance and capabilities are on-par with databases

What is a Data Warehouse?

Data Mart

- Data structured for query and reporting

Data Warehouse

- What you get if you create data marts for every system/department, then combine them together into one huge structure

Data Warehouse

- Multiple sources
- Cleansed and processed
- Organized
- Summarized

More information

www.pentaho.com/hadoop

contact: hadoop@pentaho.com