



Data Analytics in the Cloud

Tom Plunkett

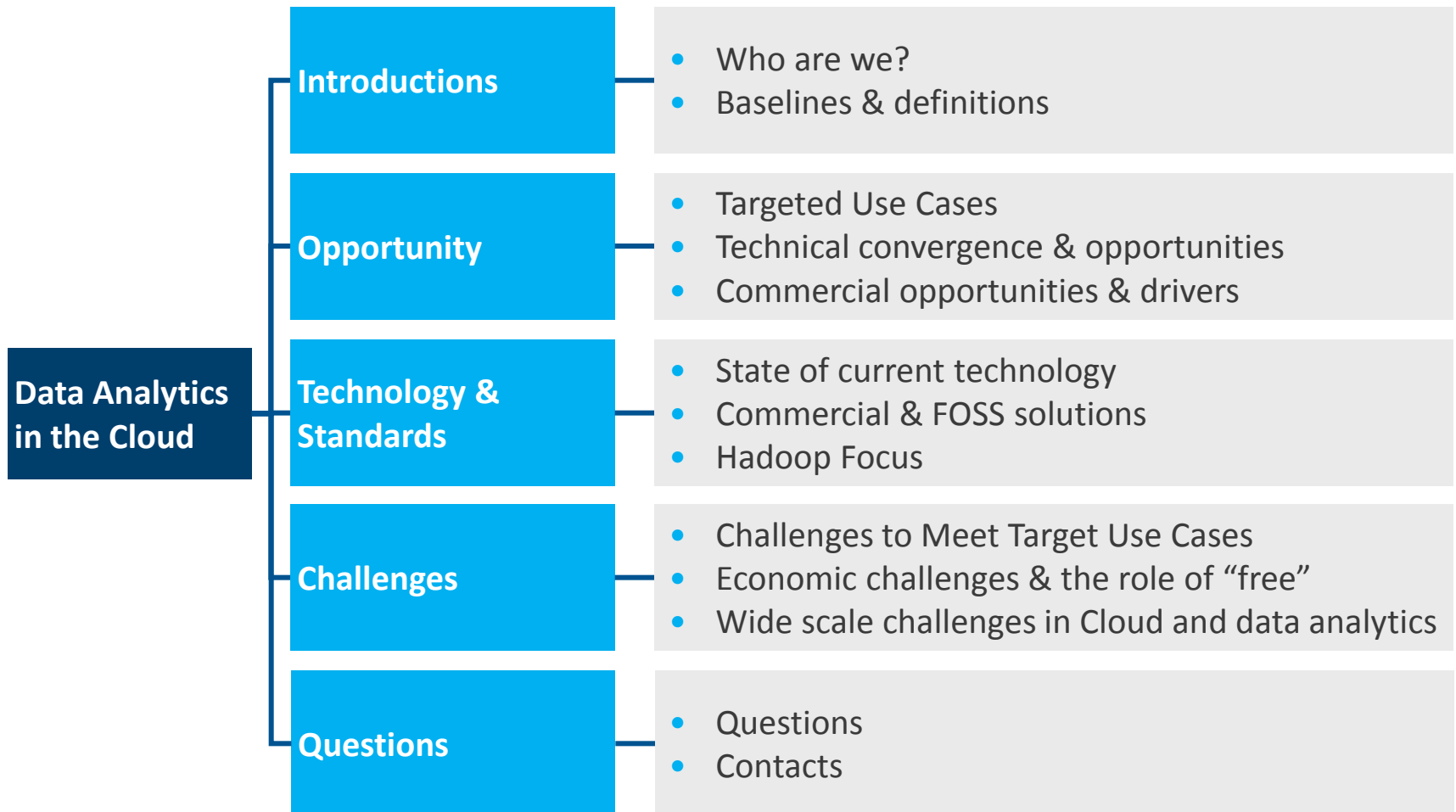
TomPlunkett@vt.edu

First International Cloud Symposium

Rotterdam

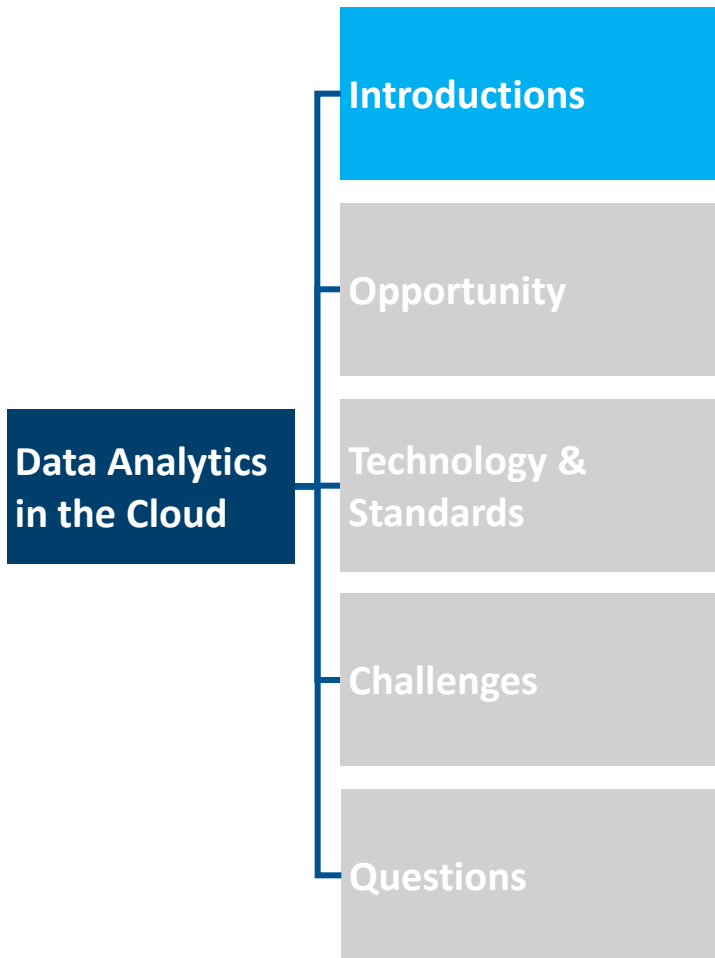
October 22-23, 2009

Overview



Cloud Data Analytics: Introductions

Introductions
Opportunity
Technology & Standards
Challenges
Questions



Tom Plunkett

Introductions
Opportunity
Technology & Standards
Challenges
Questions

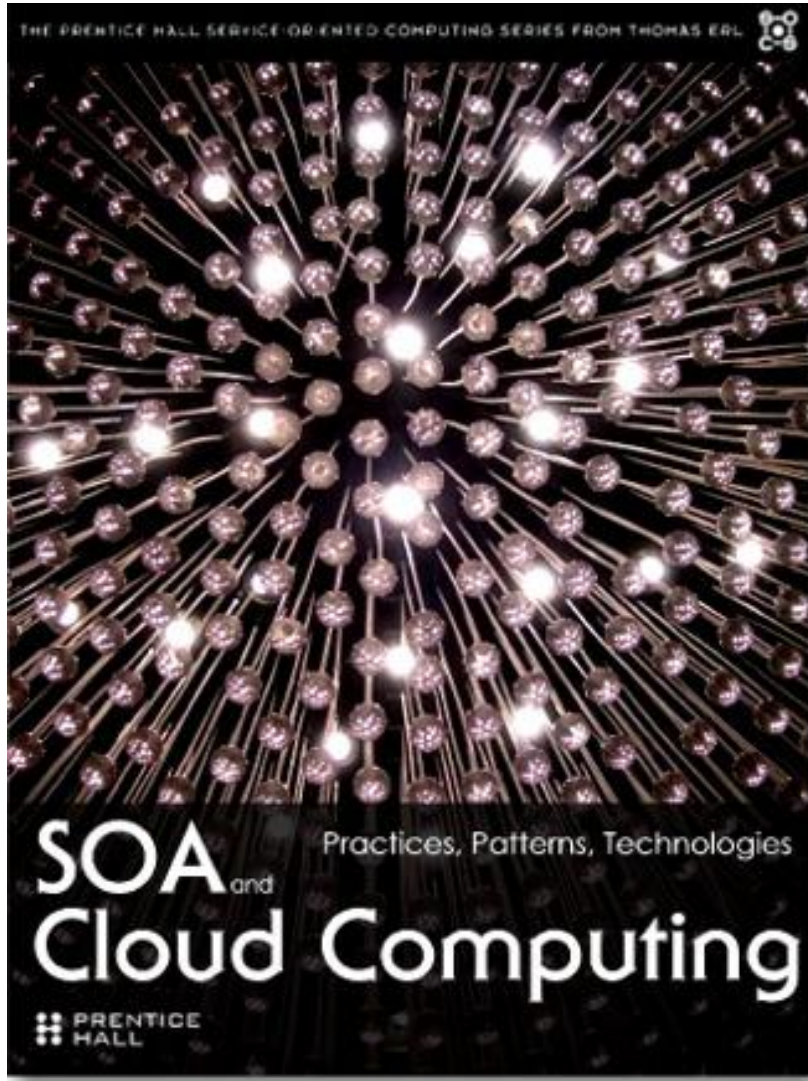
Data Analytics
in the Cloud

Cloud Computing, Java, and SOA Certifications

Extensive U.S. Federal Government Experience

Patents

Teach Cloud Computing, Java, and SOA



2Q 2010

Toufic Boubez
Thomas Erl
Nitin Gandhi
Tom Plunkett
Herbjorn Wilhelmsen

- Introductions
- Opportunity
- Technology & Standards
- Challenges
- Questions

Draft NIST Definition of Cloud Computing

A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction

Essential Characteristics	Delivery Models	Deployment Models
<ul style="list-style-type: none"> • On-demand self-service • Ubiquitous network access • Location independent resource pooling • Rapid elasticity • Measured Service 	<ul style="list-style-type: none"> • Cloud Software as a Service (SaaS) • Cloud Platform as a Service (PaaS) • Cloud Infrastructure as a Service (IaaS) 	<ul style="list-style-type: none"> • Private cloud • Community cloud • Public cloud • Hybrid cloud

Source: Draft NIST Definition of Cloud Computing, 06/2009

Data Clouds & Data Grids – What's the difference?

Often Data Clouds & Data Grids are used interchangeably, we make the following distinctions

Data Grids

- Grid computing system optimized to share large amounts of distributed data
- Focus on technical capabilities
- Often combined with computational grid computing systems
- Data often moved to compute grid for use
- Often oriented towards highly structured scientific data computing applications

Data Clouds

- Focuses on perception of infinite storage, computing capacity
- Focus on cost, virtualization & flexible capacity
- Enables scale-up/scale-down economics
- Data moved rarely, locality is a key feature
- Clouds thus far focusing on column oriented, massively scalable data stores

Sources: Wikipedia & [Grossman 1]

Cloud Data Analytics: Opportunities

- Introductions
- Opportunity**
- Technology & Standards
- Challenges
- Questions

**Data Analytics
in the Cloud**



Typical Use Cases for Cloud Data Analysis

- Log Processing
- Event Detection
- Fraud Analysis
- Relationship Maps
- Relevance Ranking
- Trend Analysis
- Unstructured Data Analysis

Use Case: Cloud Data Analytical Tools for Intelligence Community Field Analyst

Problem Statement: Analytical Tools Obsolete On Deployment, field analysts need timely, configurable data analytics. How does cloud based DA meet the needs of IC analysts

Customer Problem

- Traditional business intelligence tools require years to develop
- Field Analysts confront situations which are rapidly changing
- Petabytes of data require analysis

Cloud Analytical Tools Solution

- Recomposable Cloud Computing Data Analytical Tools
 - Apache Hadoop
 - Mashups
 - Service-Oriented Architecture

Customer Value

- Enabling field analysts to quickly build the analytical tool they need to analyze petabytes of data

Use Case: Cloud Data Scientific APIs for Scientific Researchers

Problem Statement: Scientific Researchers need to analyze petabytes of data . How does cloud based data analysis meet the needs of scientific researchers?

Customer Problem

- Traditional scientific APIs required years to develop
- Existing Data Analytical tools designed for other business problems
- Petabytes of data require analysis

Cloud Analytical Tools Solution

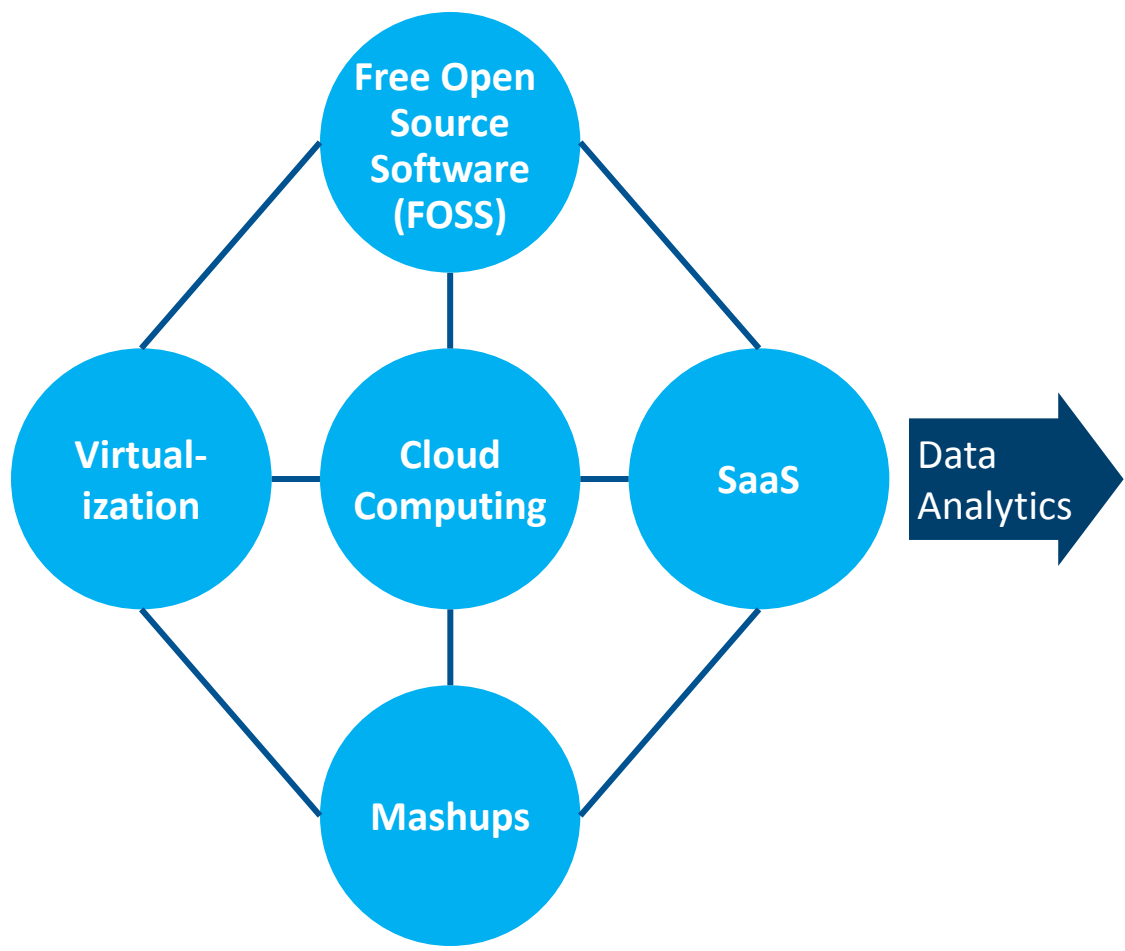
- Cloud Computing Data Analytical Tools
 - Apache Hadoop
 - Apache Hive
 - Apache Pig

Customer Value

- Enabling Researchers to quickly build the analytical tool they need to analyze petabytes of data

Why the “Buzzword” Soup? Convergence of Capabilities

Introduction
Opportunity
Technology & Standards
Challenges
Summary



Convergence of capabilities
New opportunities in breadth and depth of DA services

- **Big Data:** Cloud disk and data storage engines make petabyte environments available to new clients
- **Value Based Billing:** Heavy use of FOSS in the cloud reduces costs directly & indirectly
- **Capacity Scaling:** Scaling up/down of capacity in pay-go fashion makes DA available to wider audience
- **Composable UI's:** Capability to assemble DA results into various interfaces

Early Data Analytic Cloud Consumers/Providers

Cloud DA Opportunities

Profile	Types	Example Companies	Services
Internet Scale Service Providers	Big Internet Companies	• Yahoo, Amazon – can build DA on inf.	Services
	SaaS Companies	• Force.com – DA & Warehousing to SBA's	
	Social Platforms	• Facebook – sell DA access to anon. user info	
Large data-centric Traditional Co's	Insurers	• BCBS – private clouds across consortium	
	Healthcare & Biotech	• Kaiser Permanente – common DA services	
	Rating Agencies	• S & P – open DA cloud to customers	
Government Organizations	Intelligence Community	• CIA –private org-wide Cloud	Services
	Defense Managed Services	• DISA -- offer DA to .mil clients	
	Healthcare	• SSA – offer DA to fraud prevention analysts	
DAaaS Providers	DAaaS Infrastructure	• Cloudera –managed Hadoop instances	Services
	SMB DAaaS Provider	• ?? – managed DAaaS, simplified, low cost	

Data Analytics in the Cloud: Technology & Standards

	Introductions
	Opportunity
Data Analytics in the Cloud	Technology & Standards
	Challenges
	Questions

Data Analytics
in the Cloud

Introductions

Opportunity

Technology &
Standards

Challenges

Questions

Google MapReduce

Algorithm for computing distributed problems using a divide and conquer approach with a cluster of nodes

Master node Maps input into smaller sub-problems and distributes the work to the cluster. A worker node may further map the work for a further cluster of nodes. The worker nodes then process the smaller problems, and return the answers back to the master node

Master node then Reduces the set of answers into the answer to the original problem

Apache Hadoop

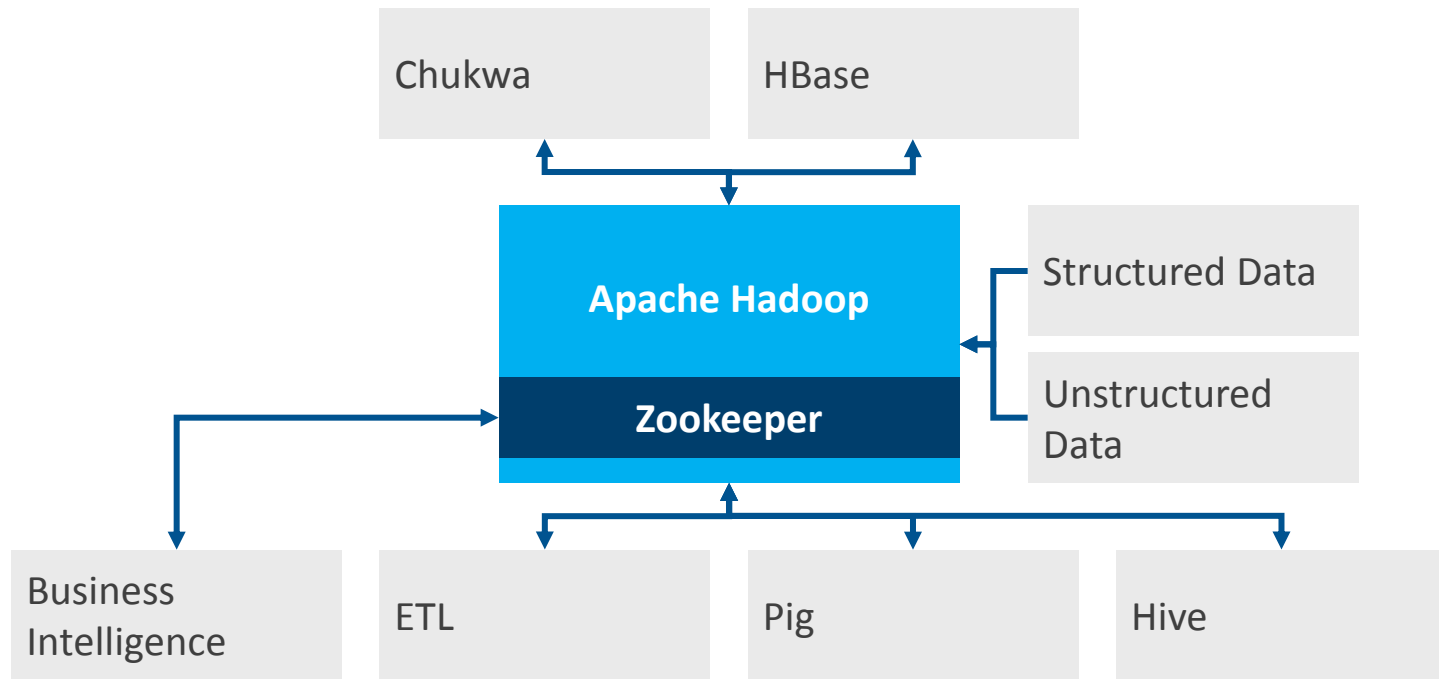
Open Source implementation of the MapReduce algorithms

Hadoop can store and process petabytes of data

Subprojects include HBase, Chukwa, Hive, Pig, and ZooKeeper

Yahoo (more than 100,000 CPUs in >25,000 computers running Hadoop) and other companies make extensive use of Hadoop

As-Is Hadoop Simplified Reference Architecture



Apache Hadoop Sub-projects

Hadoop Sub-projects

Capabilities

Example Companies

Chukwa

- Data collection system for monitoring and analyzing large distributed systems

- Yahoo

HBase

- Similar to Google's BigTable
- Distributed database for structured data
- Multi-dimensional sorted map

- Yahoo

Hive

- Data warehouse infrastructure for large datasets
- Hive QL query language

- Facebook

Pig

- High-level language for data analysis
- Compiler for Map-Reduce programs

- Yahoo

Zookeeper

- Configuration, Naming, Distributed Synchronization, and group services

- Yahoo

Hello World Mapper Example

```

public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text,
    IntWritable>
{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable>
output, Reporter reporter) throws IOException
    {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens())
        {
            word.set(tokenizer.nextToken());
            output.collect(word, one);
        }
    }
}

```

Hello World Reducer Example

```
public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text,
    IntWritable>
{
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable>
        output, Reporter reporter) throws IOException
    {
        int sum = 0;
        while (values.hasNext())
        {
            sum += values.next().get();
        }
        output.collect(key, new IntWritable(sum));
    }
}
```

Hello World Application Example

```

public static void main(String[] args) throws Exception {
    JobConf conf = new JobConf(WordCount.class);
    conf.setJobName("wordcount");

    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(IntWritable.class);

    conf.setMapperClass(Map.class);
    conf.setCombinerClass(Reduce.class);
    conf.setReducerClass(Reduce.class);

    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);

    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));

    JobClient.runJob(conf); }

```

Hello World Execution

Sample Input Files

Hello World Bye World

Hello Hadoop Goodbye Hadoop

Output:

Bye 1

Goodbye 1

Hadoop 2

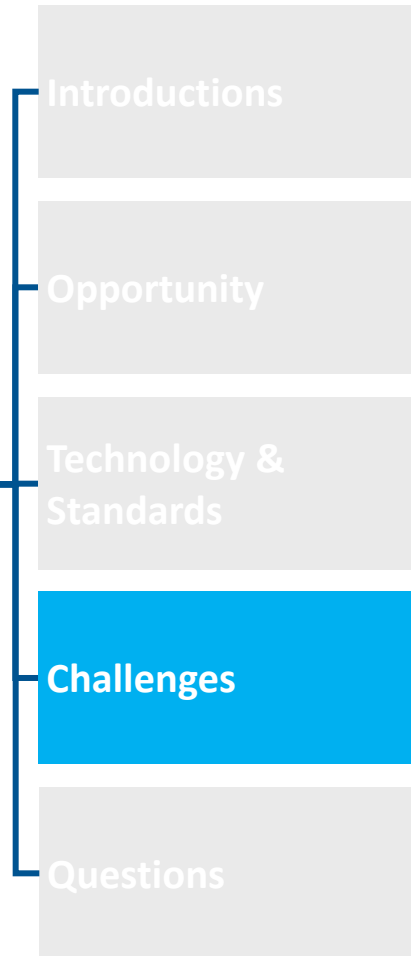
Hello 2

World 2

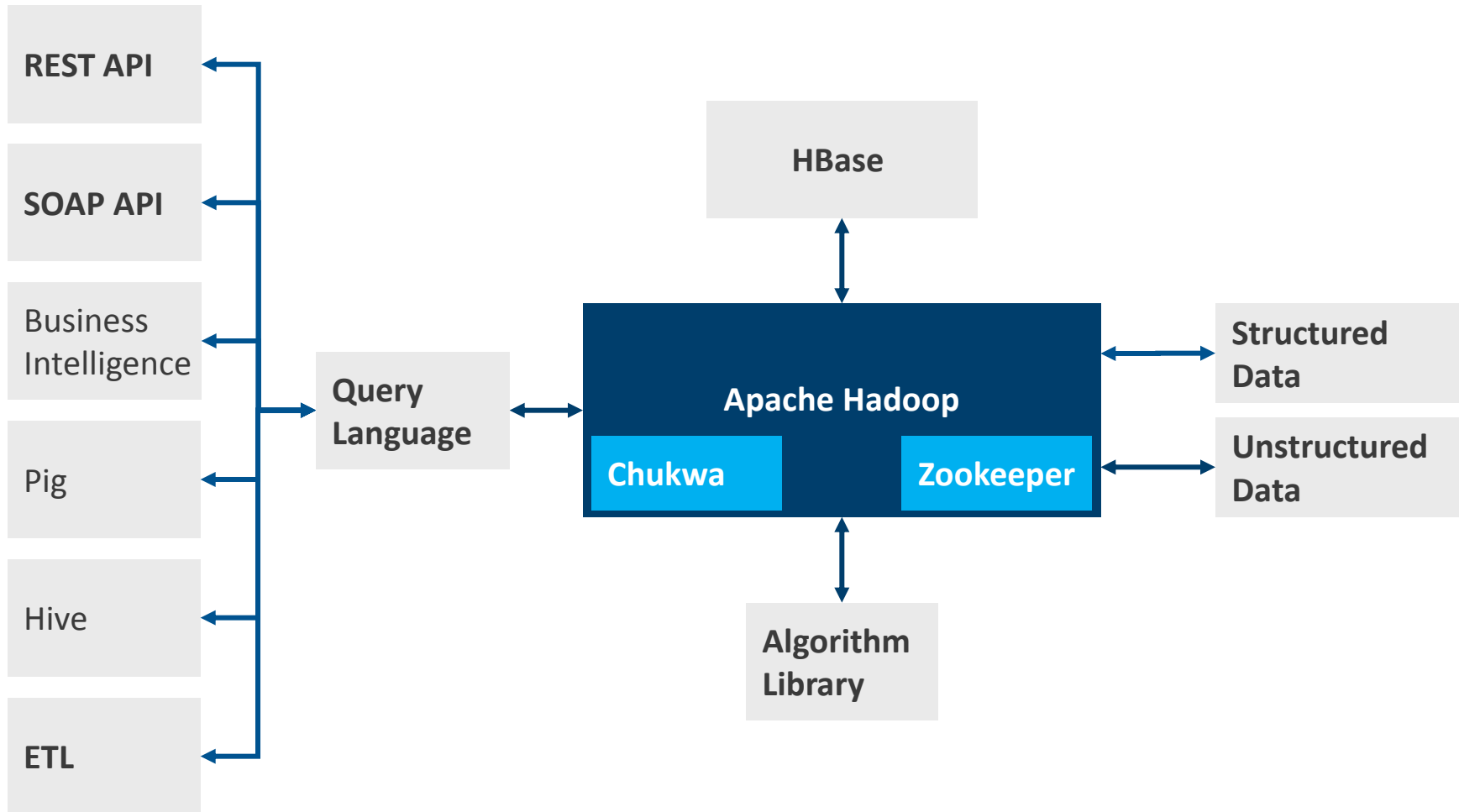
Data Analytics in the Cloud: Challenges

- Introductions
- Opportunity
- Technology & Standards
- Challenges**
- Questions

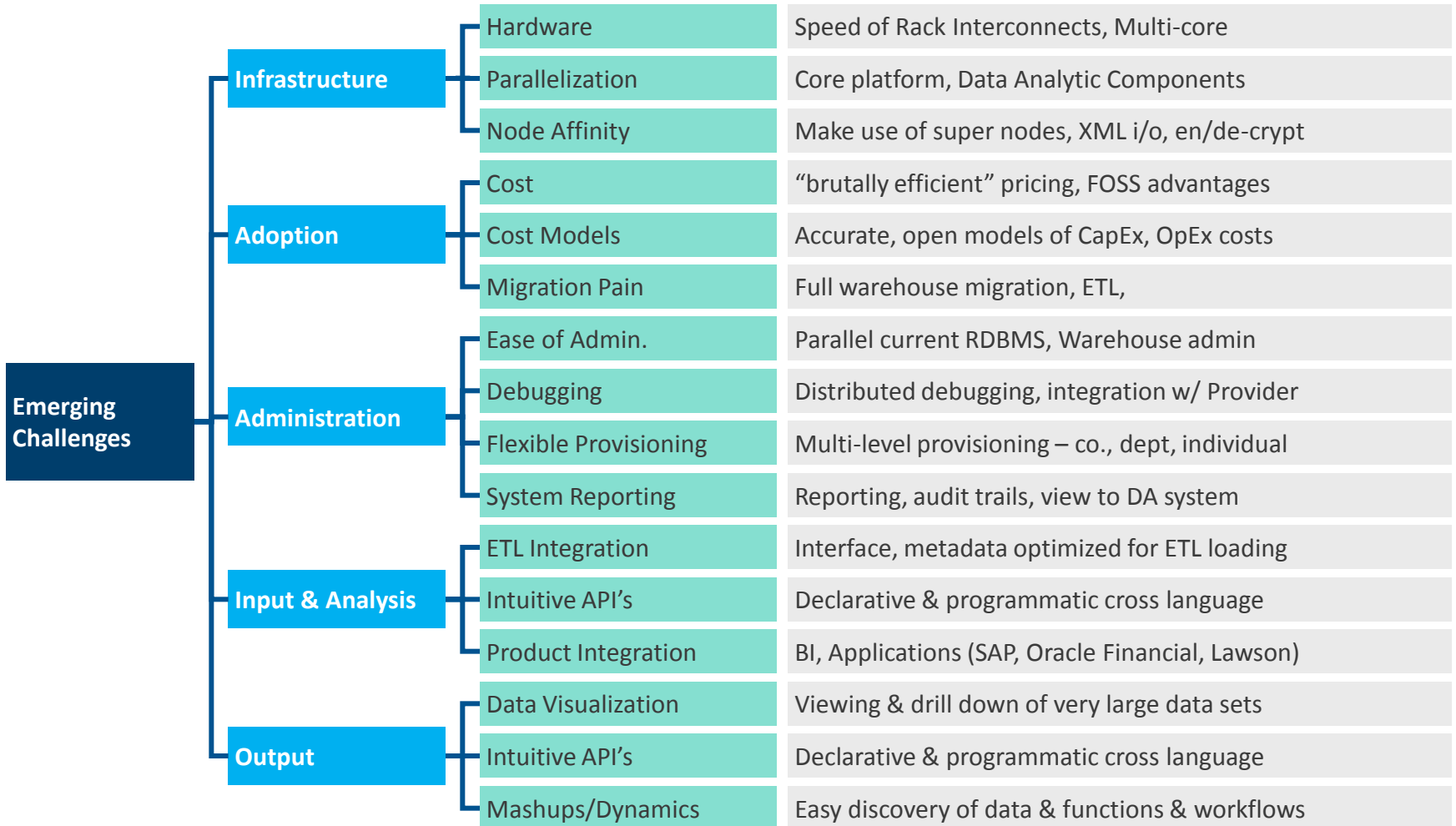
Data Analytics in the Cloud



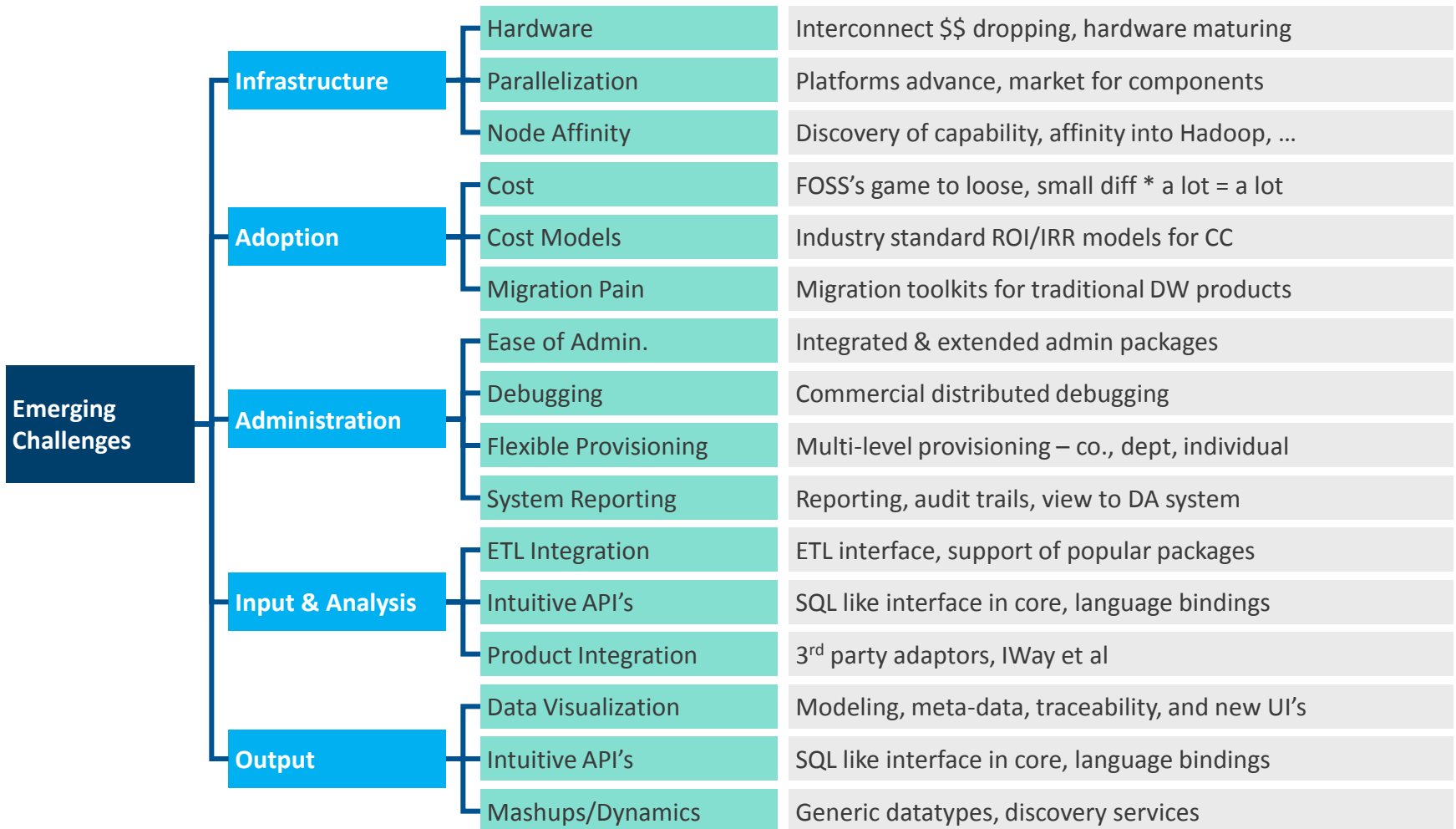
To-Be Simplified Hadoop Architecture



Key Challenges



Solutions: Projected & In-Progress



Resources

<http://hadoop.apache.org/>

Books

Jason Venner, Pro Hadoop

Tom White, Hadoop: The Definitive Guide

Data Analytics in the Cloud: Questions

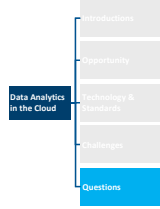
- Introductions
- Opportunity
- Technology & Standards
- Challenges
- Questions

Data Analytics in the Cloud



Questions?

TomPlunkett@vt.edu



A vertical table of contents on the right side of the slide. It consists of a dark blue header 'Data Analytics in the Cloud' and five items: 'Introduction', 'Opportunity', 'Technology & Standards', 'Challenges', and 'Questions'. The 'Questions' item is highlighted with a blue background, while the others are light gray.

Data Analytics in the Cloud	
Introduction	
Opportunity	
Technology & Standards	
Challenges	
Questions	