2013 Esri International User Conference
July 8–12, 2013 | San Diego, California

Technical Workshop

Big Data: Using ArcGIS with Apache Hadoop

David Kaiser

Erik Hoel

Offering 1330

In this technical workshop

- This presentation is for anyone that uses ArcGIS and is interested in analyzing large amounts of data

- We will cover:
  - Big Data overview
  - The Hadoop platform
  - How Esri's *GIS Tools for Hadoop* enables developers to process spatial data on Hadoop
  - How ArcGIS can leverage these custom Hadoop applications for GIS analysis

Big Data Overview

Big Data

- Within ArcGIS, Geoprocessing was enhanced at 10.1 SP1 to support 64-bit address spaces
  - This is sufficient to handle traditional large GIS datasets
- However, this solution may run into problems when confronted with datasets of a size that are colloquially referred to as Big Data
  - Internet scale datasets

Age of Data Ubiquity

- Data is now central to our existence – both for corporations and individuals
- Nimble, thin, data-centric apps exploiting massive data sets generated by both enterprises and consumers
- Hardware era: 20 – 30 years
- Software era: 20 – 30 years
- Data era: ?

Age of Data Ubiquity

- Data is now central to our existence – both for corporations and individuals
- Nimble, thi
  data sets g
  consumers

> *"The Internet has caused a Cambrian explosion of new life forms, new applications, new data-centric APIs, … literally thousands and thousands of new APIs are born every month."*
> **- Mike Hoskins**

- Hardware era: 20 – 30 years
- Software era: 20 – 30 years
- Data era: ?

Sensor data challenge

- Sensors are continuously observing our world
  - With increases in networking technology, we are moving away from pure remote sensing and moving towards direct sensing, where every sensor reports it's own data
- One researcher says: "We are instrumenting the universe", referring to what is often called *The Internet of Things*

Internet of Things

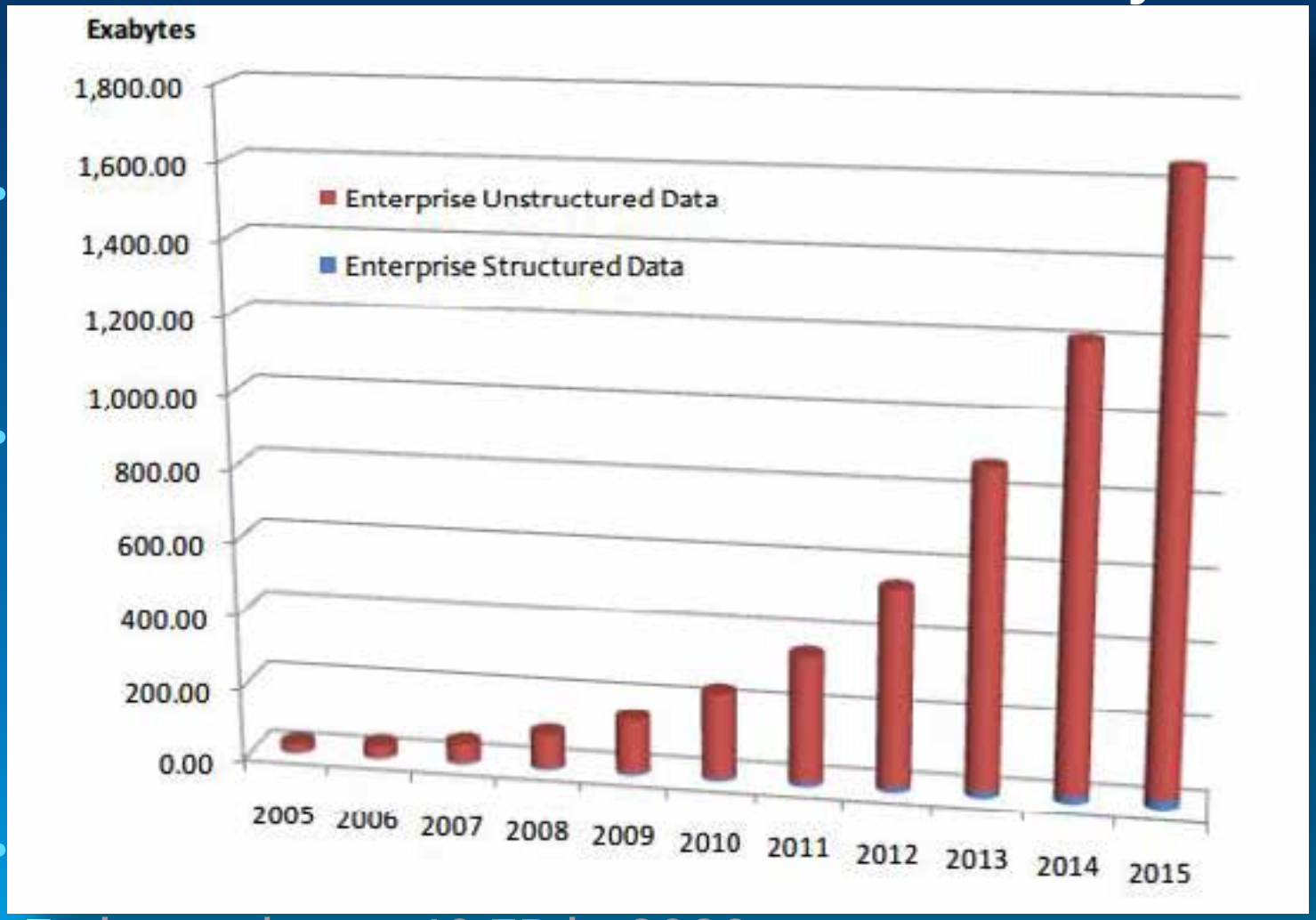
IoT - uniquely identifiable objects and their virtual representations in an Internet-like structure; the term was proposed by Kevin Ashton (MIT) in 1999

- Your old refrigerator was a dumb refrigerator
- Your new refrigerator is a smart refrigerator
  - It's a digital asset aware of itself
  - It records time, temperatures, vibrations
  - It records the electricity it's consuming
  - It's connected to the Internet as an IP device
- More realistically, commercial jets or valuable equipment
- By continuously observing things, we obtain a massive amount of valuable information

Volume of data – the rise of data anarchy

- If all 7 billion people on Earth joined Twitter and continually tweeted for one century, they would generate one zettabyte of data (billion terabytes)
- Almost double that amount, 1.8 zettabytes (1.8×10^{21}), was generated globally in 2011 (and 2.8 ZB in 2012)
- Rising 40-50% per year
- Estimated over 40 ZB in 2020

Volume of data – the rise of data anarchy



- Estimated over 40 ZB in 2020

Example

- For every hour that it runs, a commercial jet aircraft can create 20 gigabytes of operations information
- For a single journey across the Atlantic Ocean, a two-engine jet can create over 125 GB of data
- Multiply that by the more than 100,000 flights flown each day, and you get an understanding of the enormous amount of data that exists
  - E.g., 5-10 petabytes/day; 2-4 exabytes/year

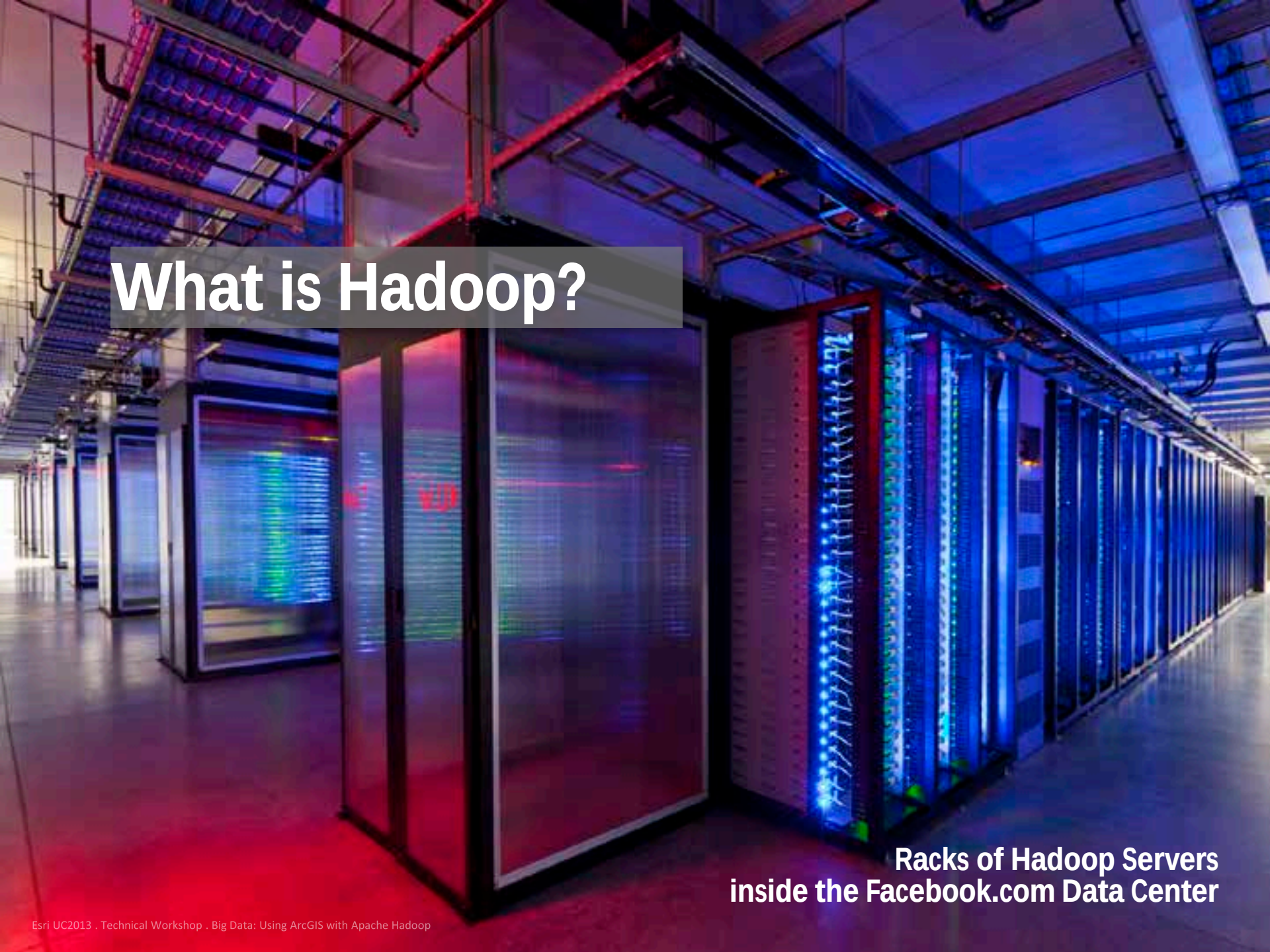
All this data has business value

- Smart Sensors
  - Electrical meters
- GPS Telemetry
  - Vehicle tracking, smartphone data collectors
- Internet Services
  - E-Commerce transactions, social media
- Monitoring Sensors
  - Heavy equipment, aircraft, etc.

Value when analyzing data at mass scale

- As observations increase in frequency
  - Each individual observation is worth less
  - ...as the set of all observations becomes more valuable
- One single metric from the jet aircraft is much less useful than the analysis of that metric against the same metric from every known flight of that aircraft over time
- *Big Data* is the accumulation and analytical processes that uses this data for business value

What is Hadoop?



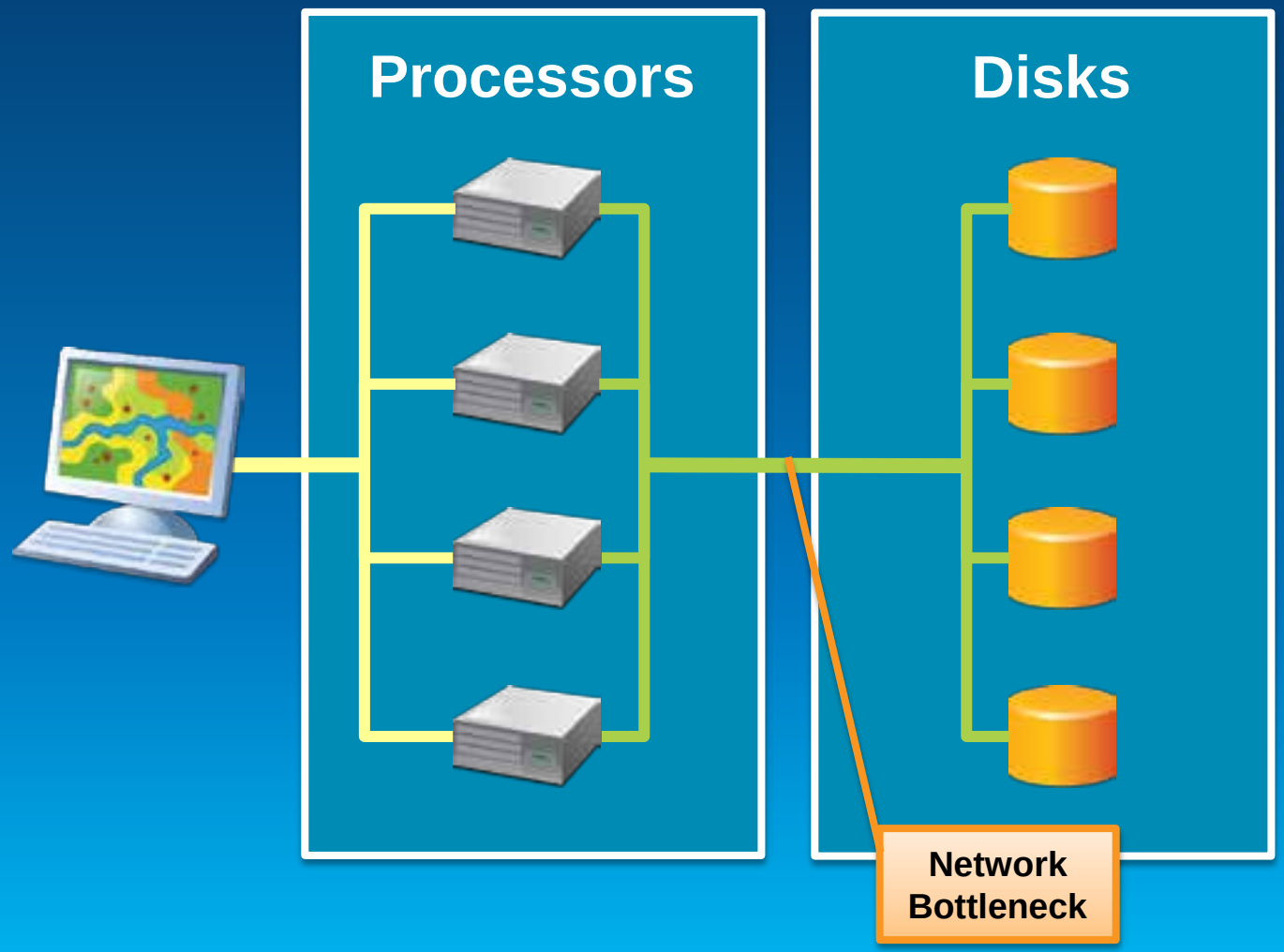
Racks of Hadoop Servers
inside the Facebook.com Data Center

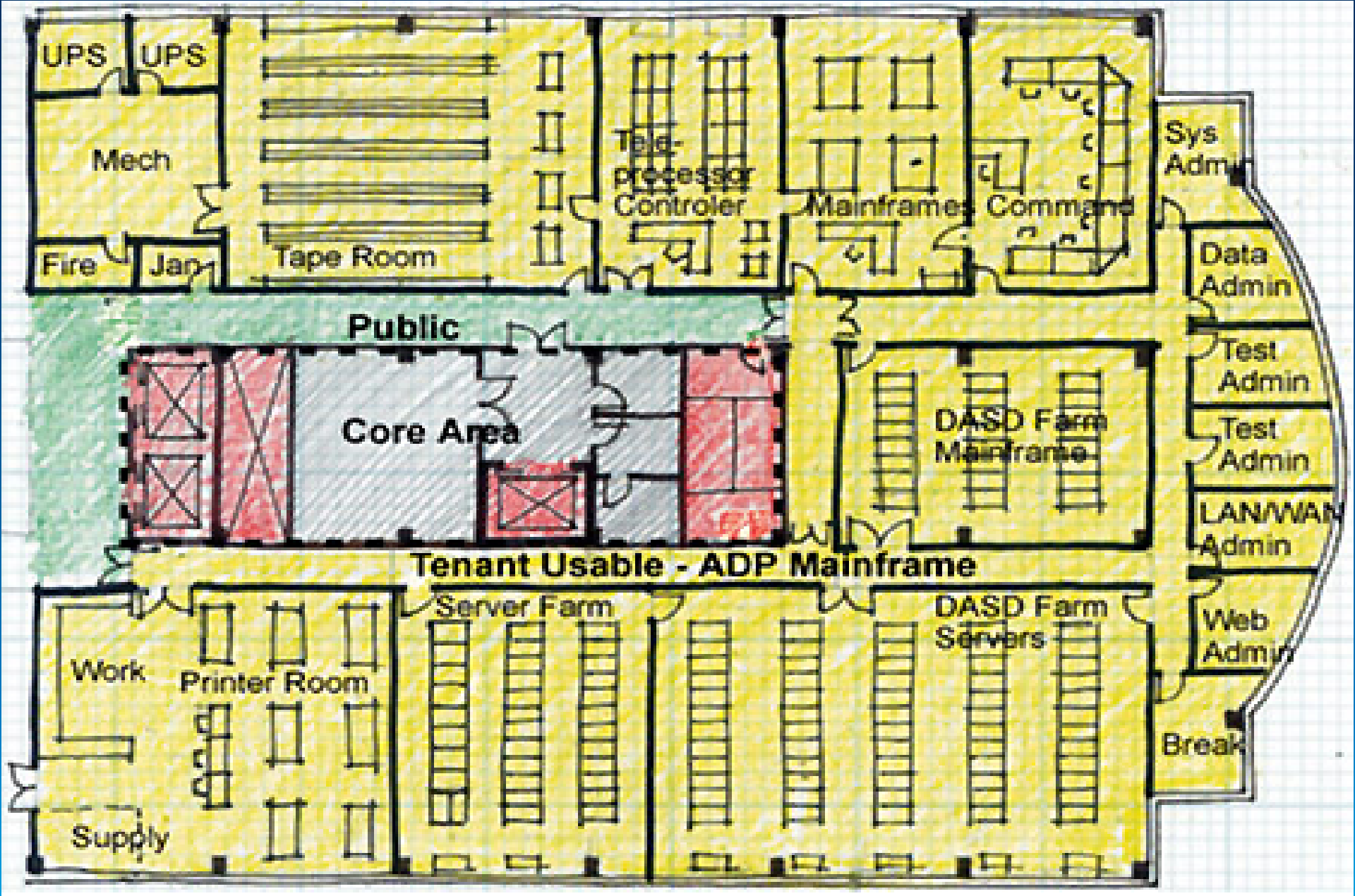
The need for distributed systems

- As data volumes have grown, individual computing systems have not scaled equivalently, e.g.,
  - Single-node databases
  - Single computers

- Big data has led to distributed systems:
  - Data is stored across a number of servers
  - Processing the data takes place in the server where the data is stored

Legacy system architecture



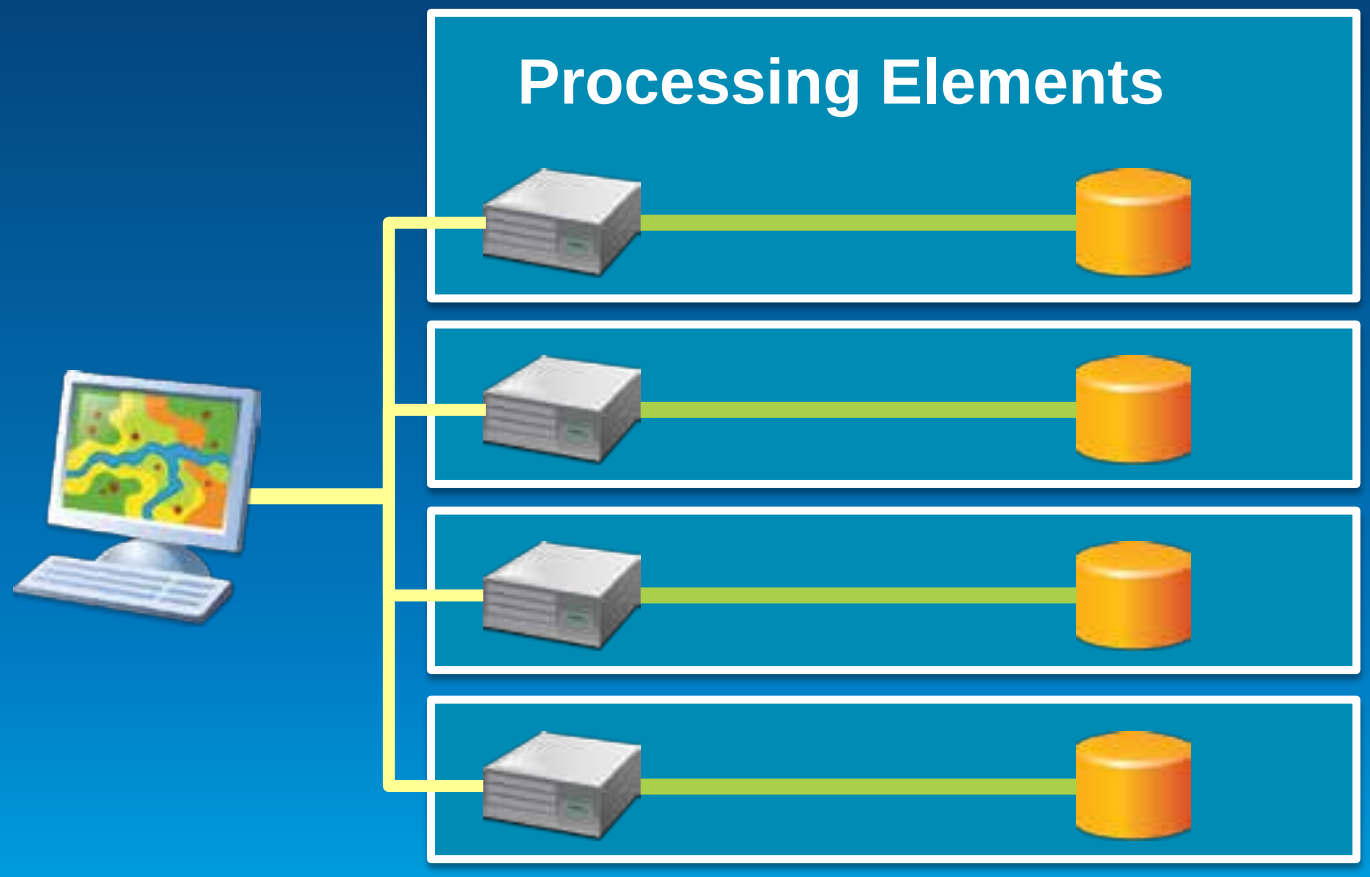


ADP Mainframe Space Type

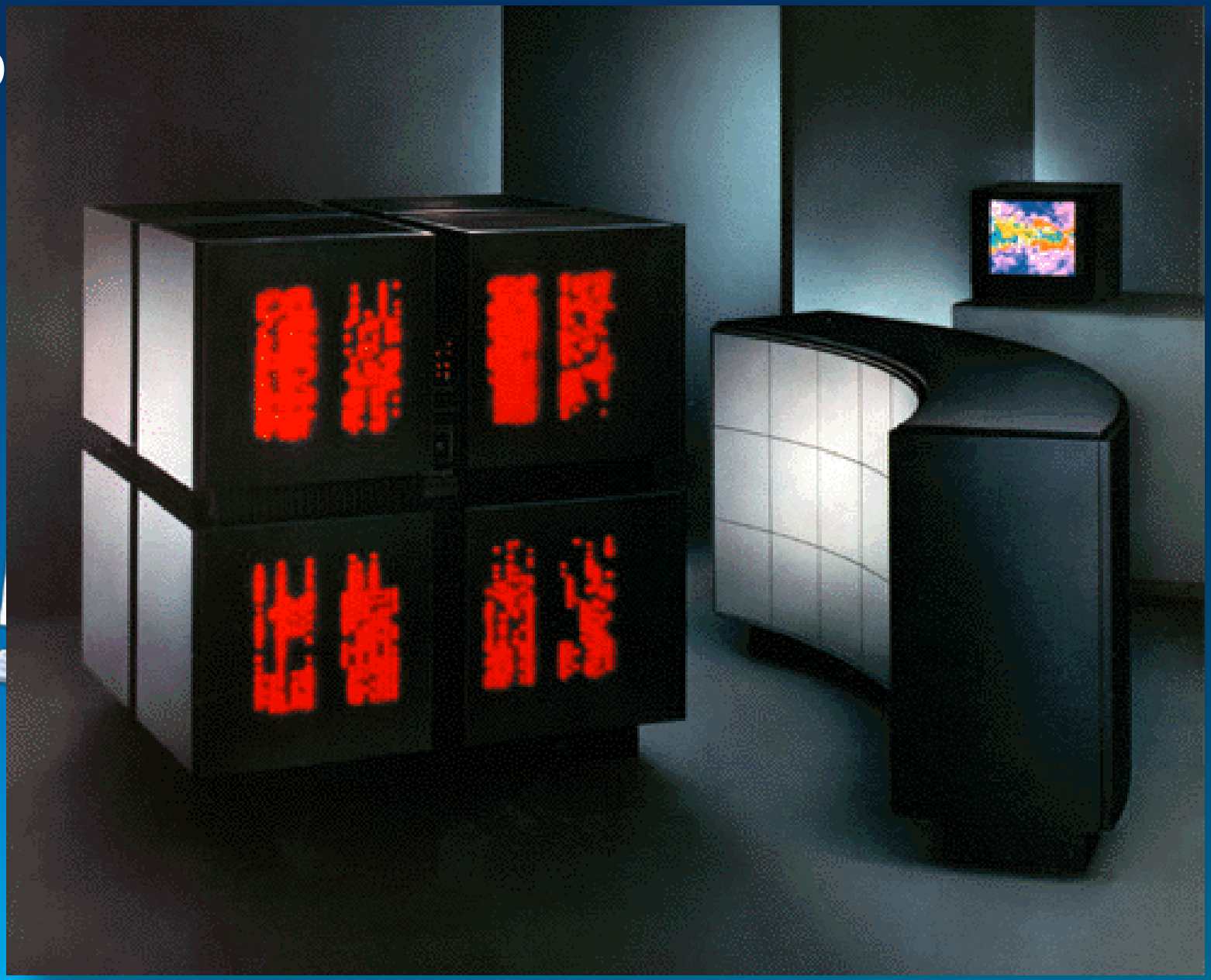
0 10'-0" 20'-0" 40'-0"

ADP Mainframe Space Type

Distributed system architecture



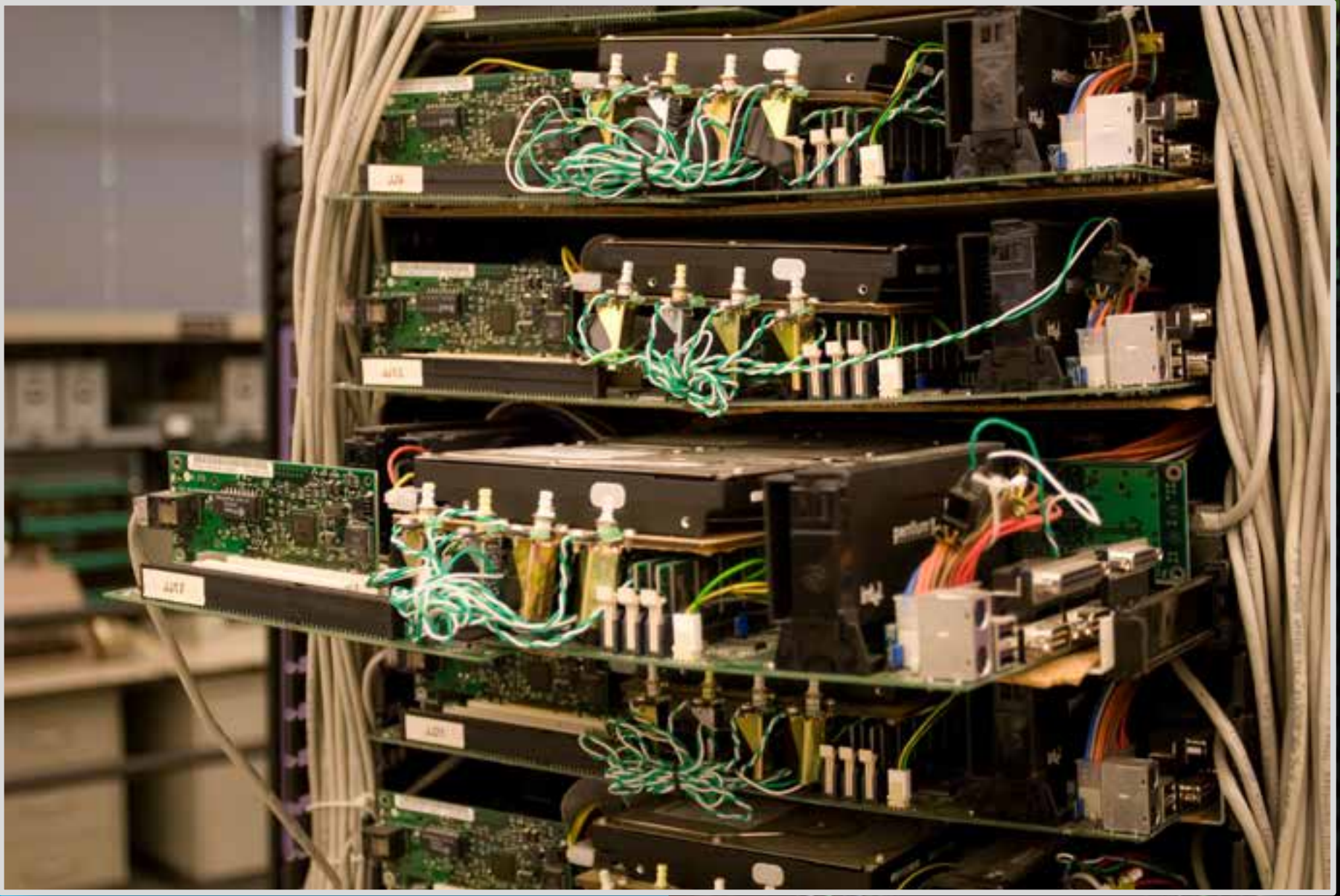
D



Until Now…

- Google implemented their enterprise on a distributed network of many nodes, fusing storage and processing into each node

- Hadoop is an open source implementation of the framework that Google has built their business around for many years



Hadoop

- Open-source data processing framework
- Provides scalable, distributed system for both storage and computation of data
  - It supports the running of applications on large clusters of commodity hardware
  - Hadoop was derived from Google's MapReduce and Google File System (GFS) scientific papers
- Platform consists of the kernel, HDFS, and MapReduce
  - Other applications have been built on Hadoop; this includes Hive, HBase, etc.

MapReduce

- A programming model for processing large data sets with a parallel distributed algorithm on a cluster
- A MapReduce program is comprised of:
  - A Map() procedure that performs filtering and comparing, and
  - A Reduce() procedure that performs a summary operation
- The MapReduce System marshals the distributed servers, runs the various tasks in parallel, manages all communications and data transfers, and provides for redundancy and failures
- MapReduce libraries have been written in many languages; Hadoop is a popular open source implementation

High Level MapReduce Walk-Through

- An instance of a MapReduce program is a job
- The job accepts arguments for data input and outputs
- The job combines
  - Functions from the MapReduce framework
    - “Splitting” large inputs into smaller pieces
    - Reading inputs and writing outputs
  - Functions that are written by the application developer
    - *Map* function, maps input values to keys
    - *Reduce* function, reduces many keys to one

High Level MapReduce Walk-Through

- An instance of a MapReduce program is a job

MapReduce Job

High Level MapReduce Walk-Through

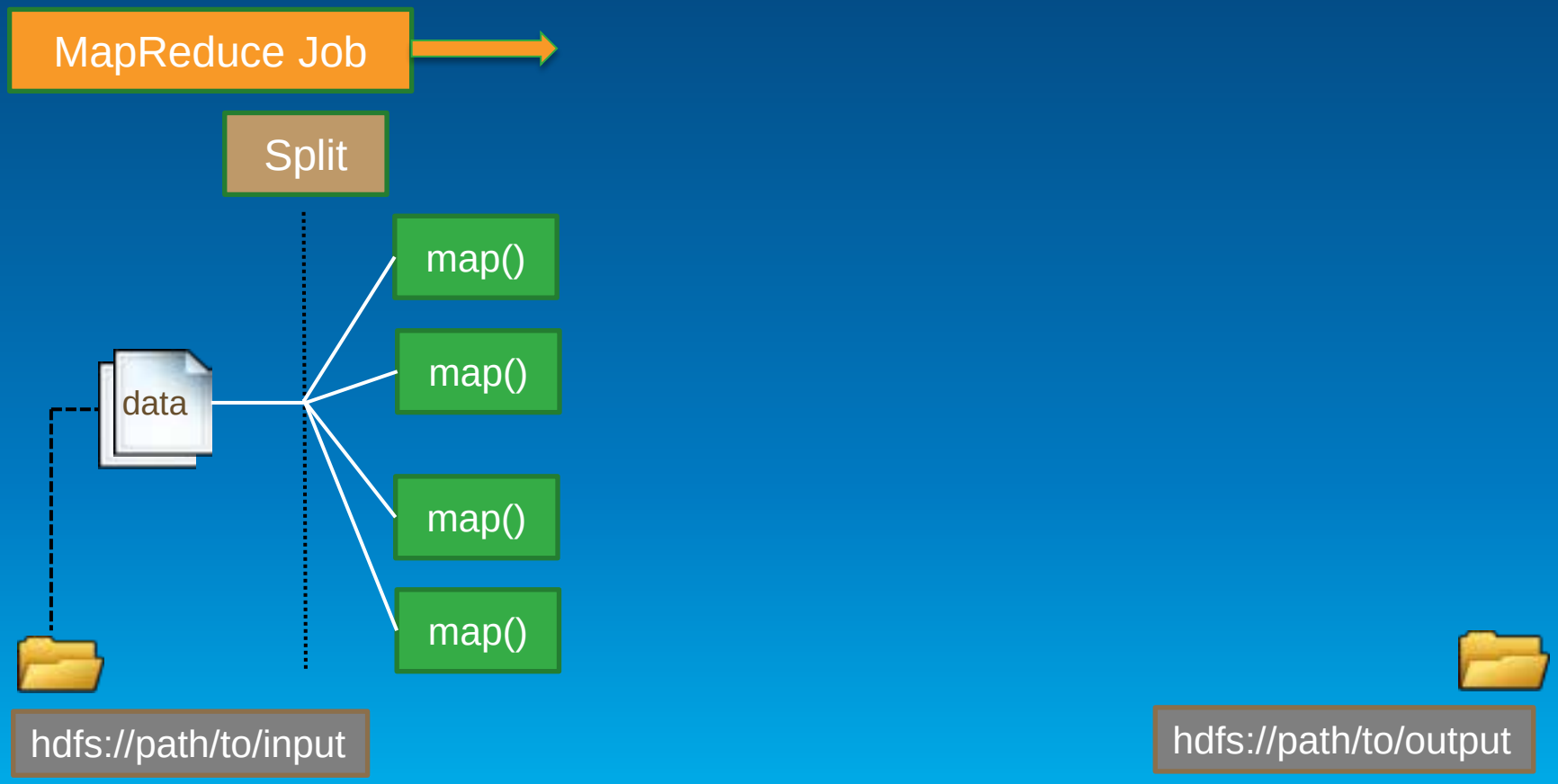
- The job accepts arguments for data input and outputs

High Level MapReduce Walk-Through

- The MapReduce framework splits the data space

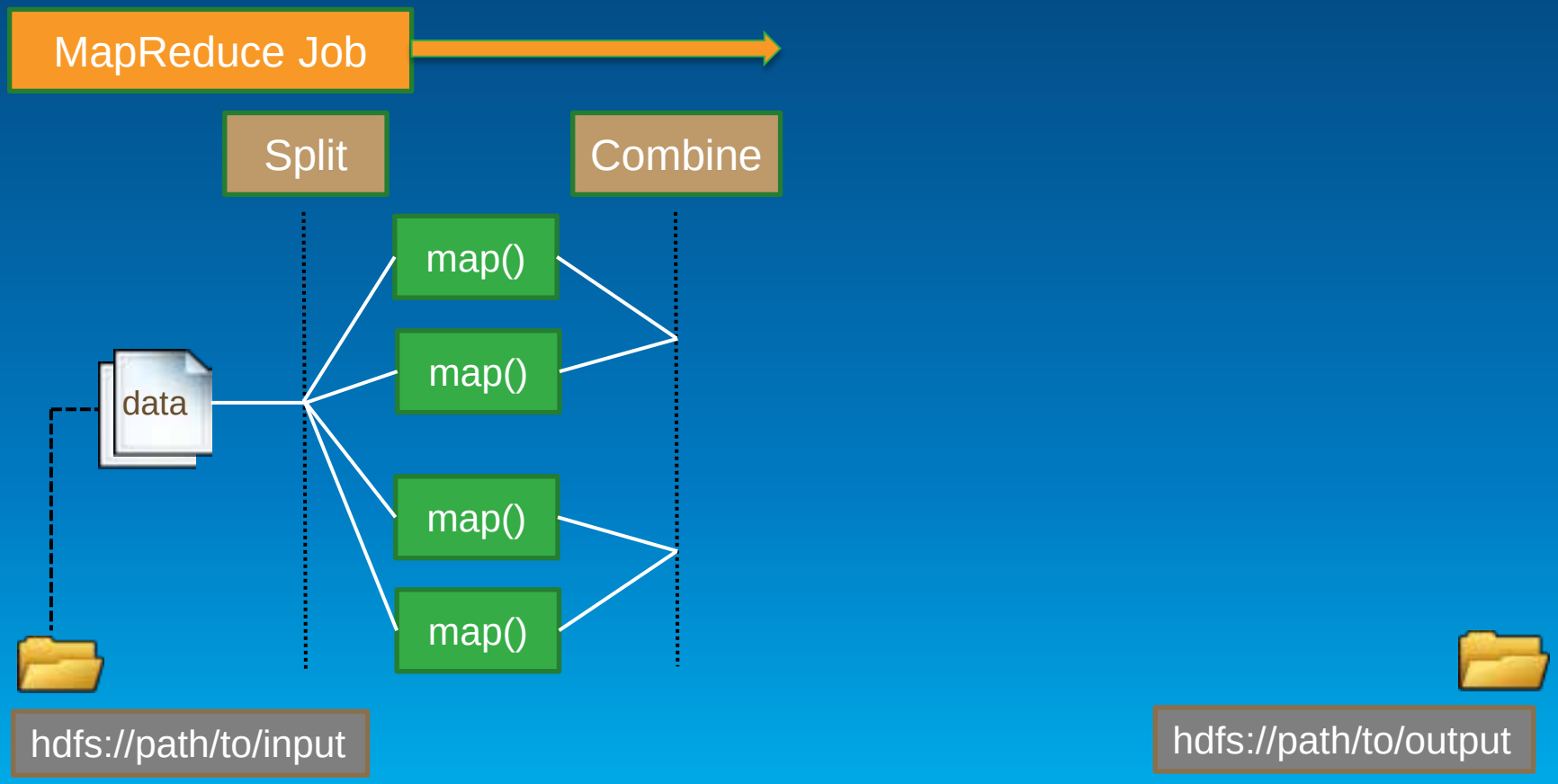
High Level MapReduce Walk-Through

- Many map() functions are run in parallel



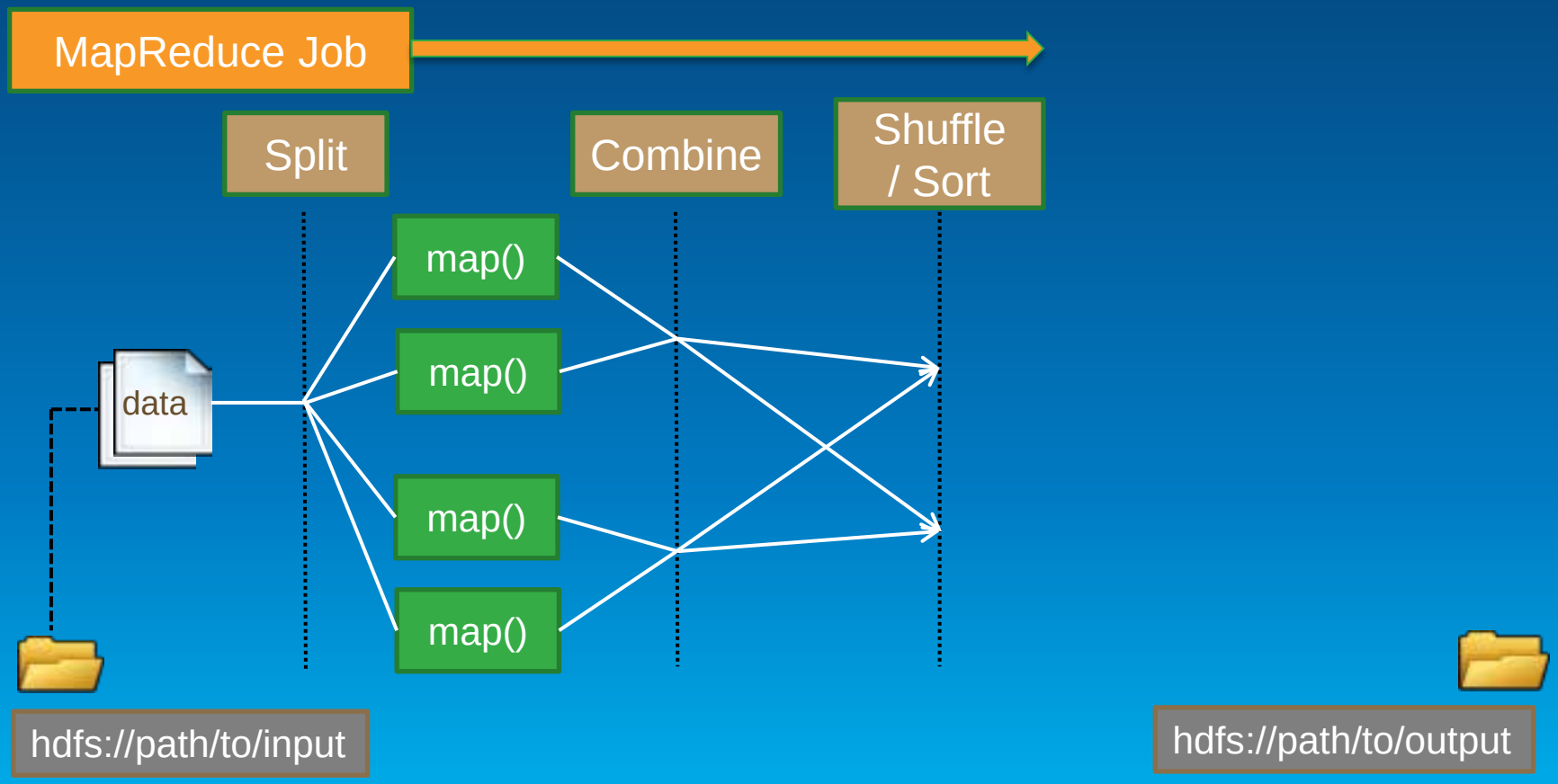
High Level MapReduce Walk-Through

- The framework runs combine to join map() outputs



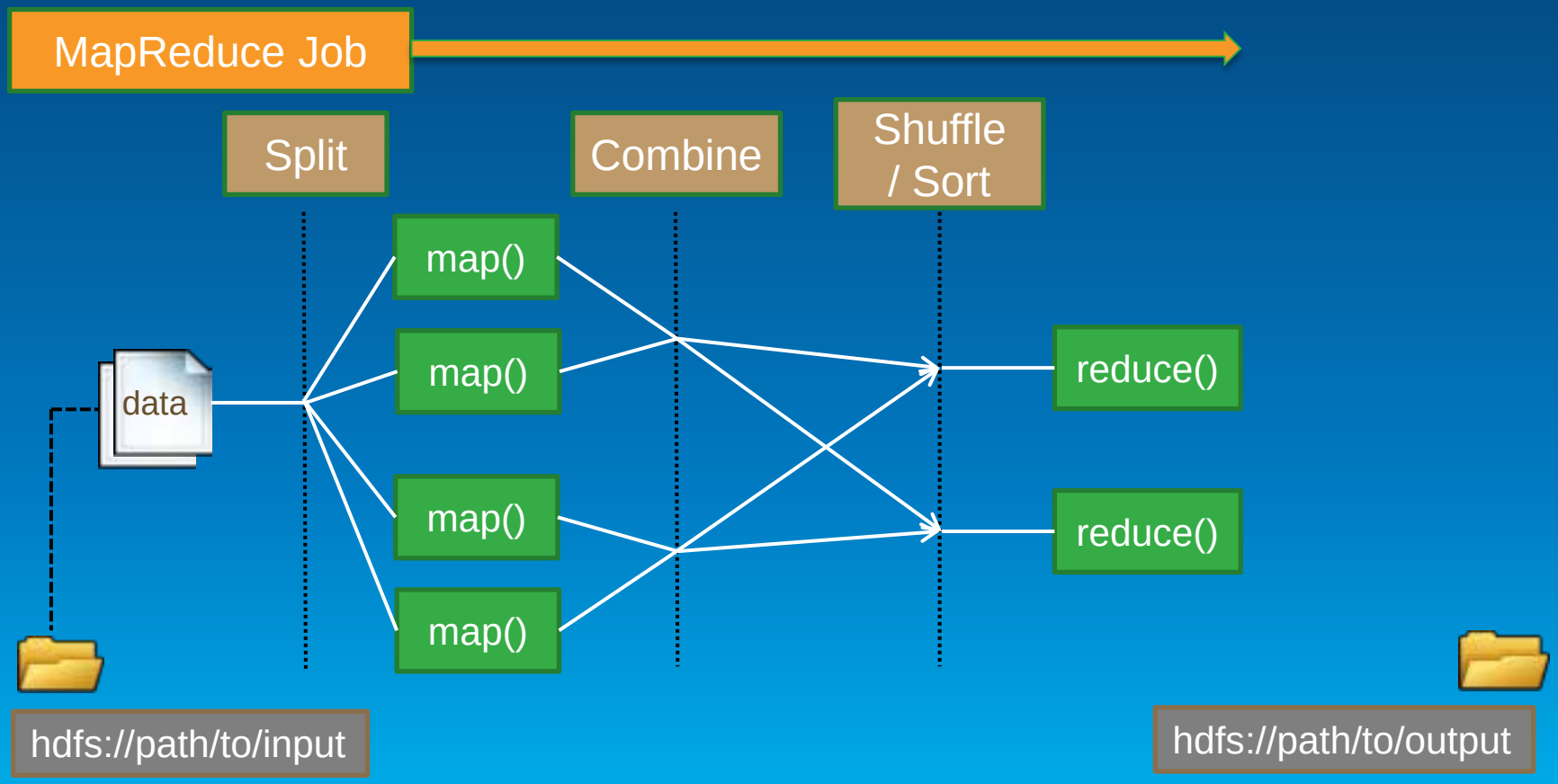
High Level MapReduce Walk-Through

- The framework performs a shuffle and sort on all data



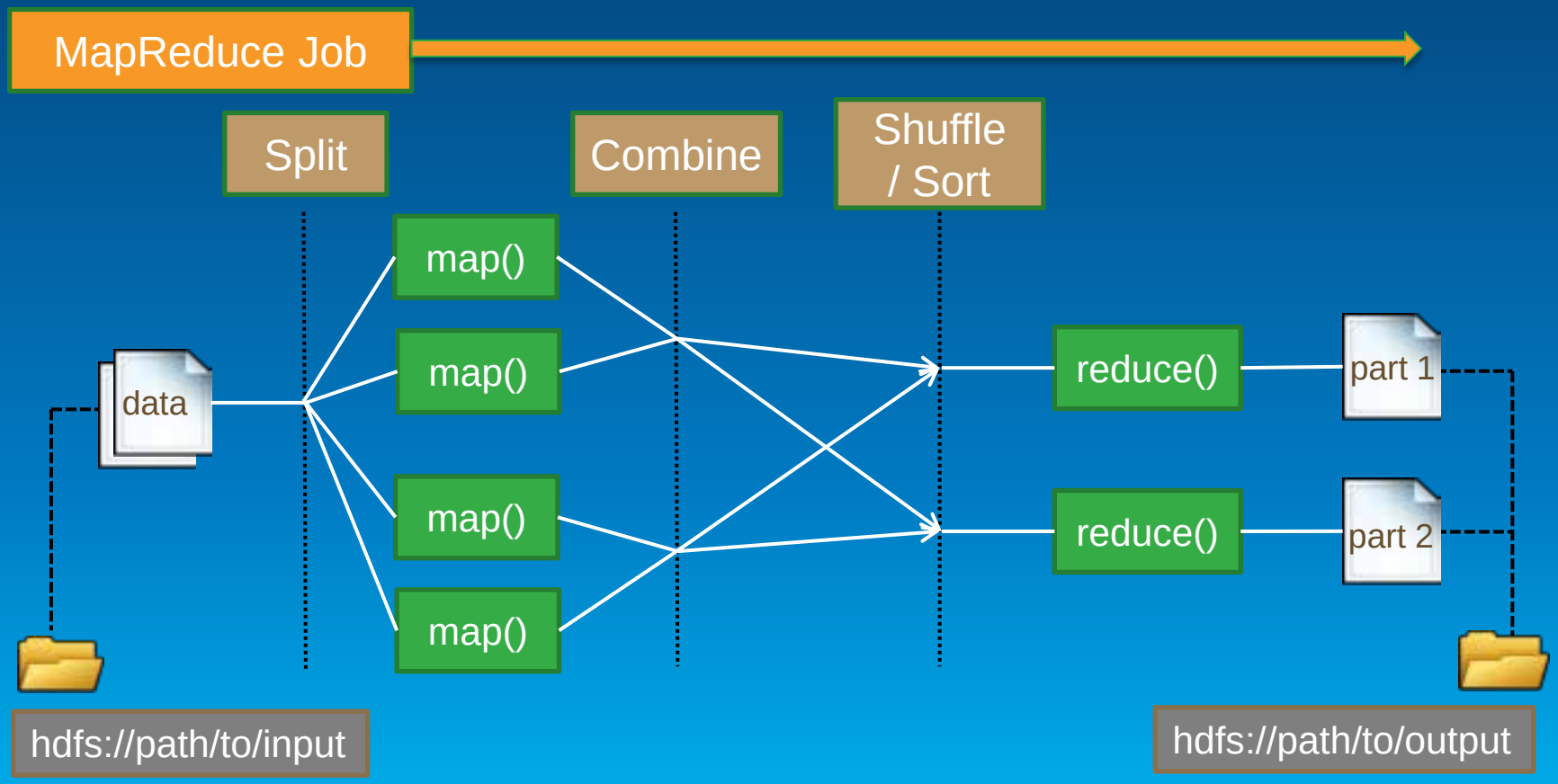
High Level MapReduce Walk-Through

- The reduce() functions work against the sorted data



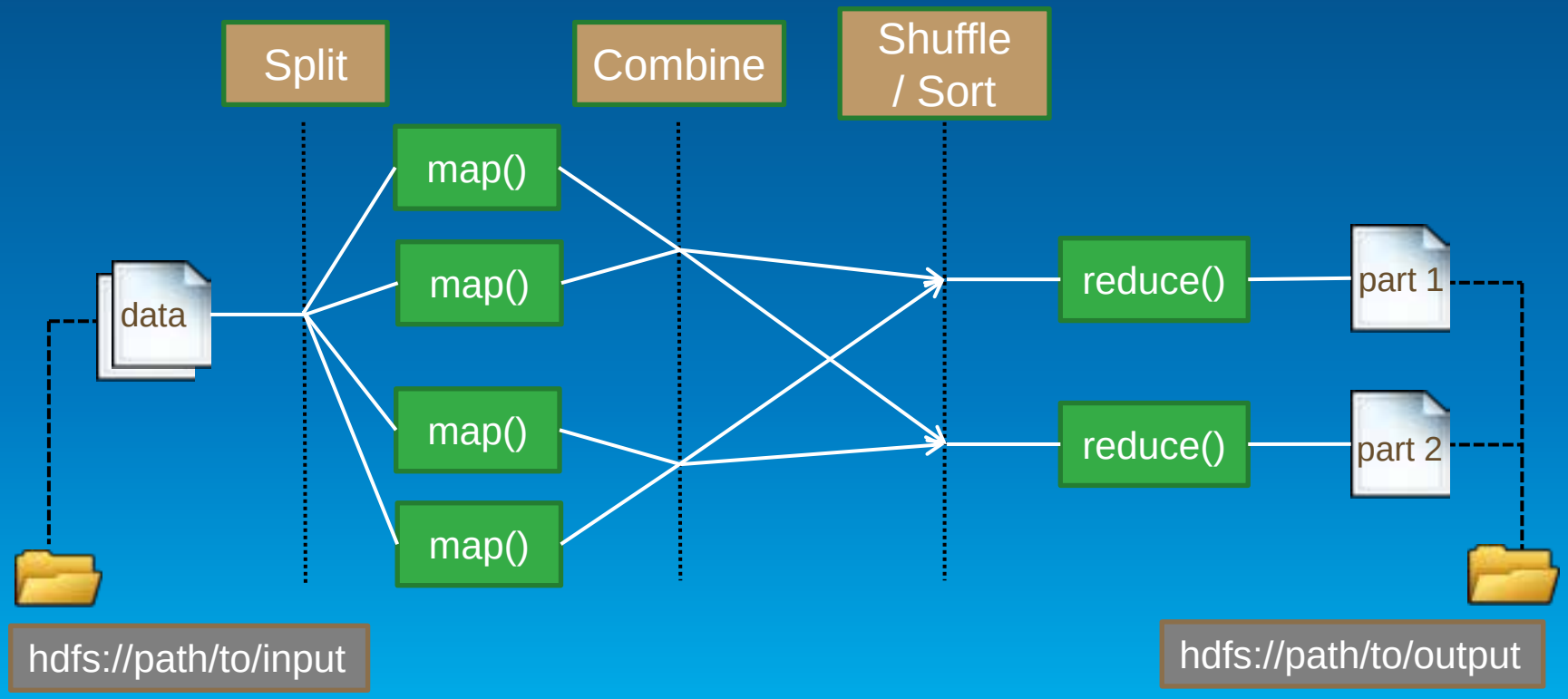
High Level MapReduce Walk-Through

- Reducers each write a part of the results to a file

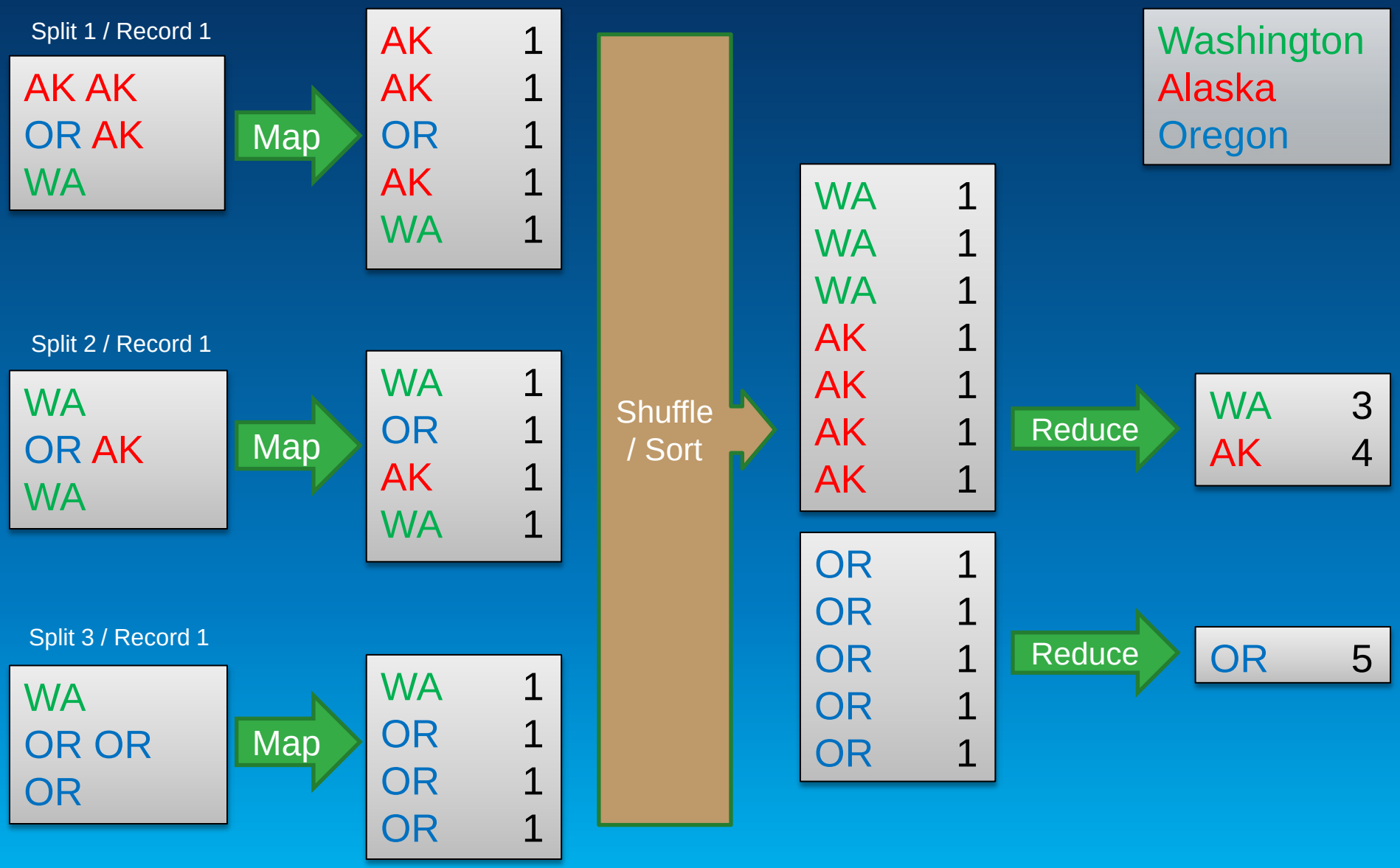


High Level MapReduce Walk-Through

- When the job has finished, ArcGIS can retrieve the results



MapReduce – Polygon Count Example



Yahoo! Hadoop Cluster in 2011

- ~10,000 servers in a number of clusters
- Largest cluster is 1,600 nodes
- Nearly 1 Petabyte of user data
- Yahoo! runs nearly 10,000 research jobs per month

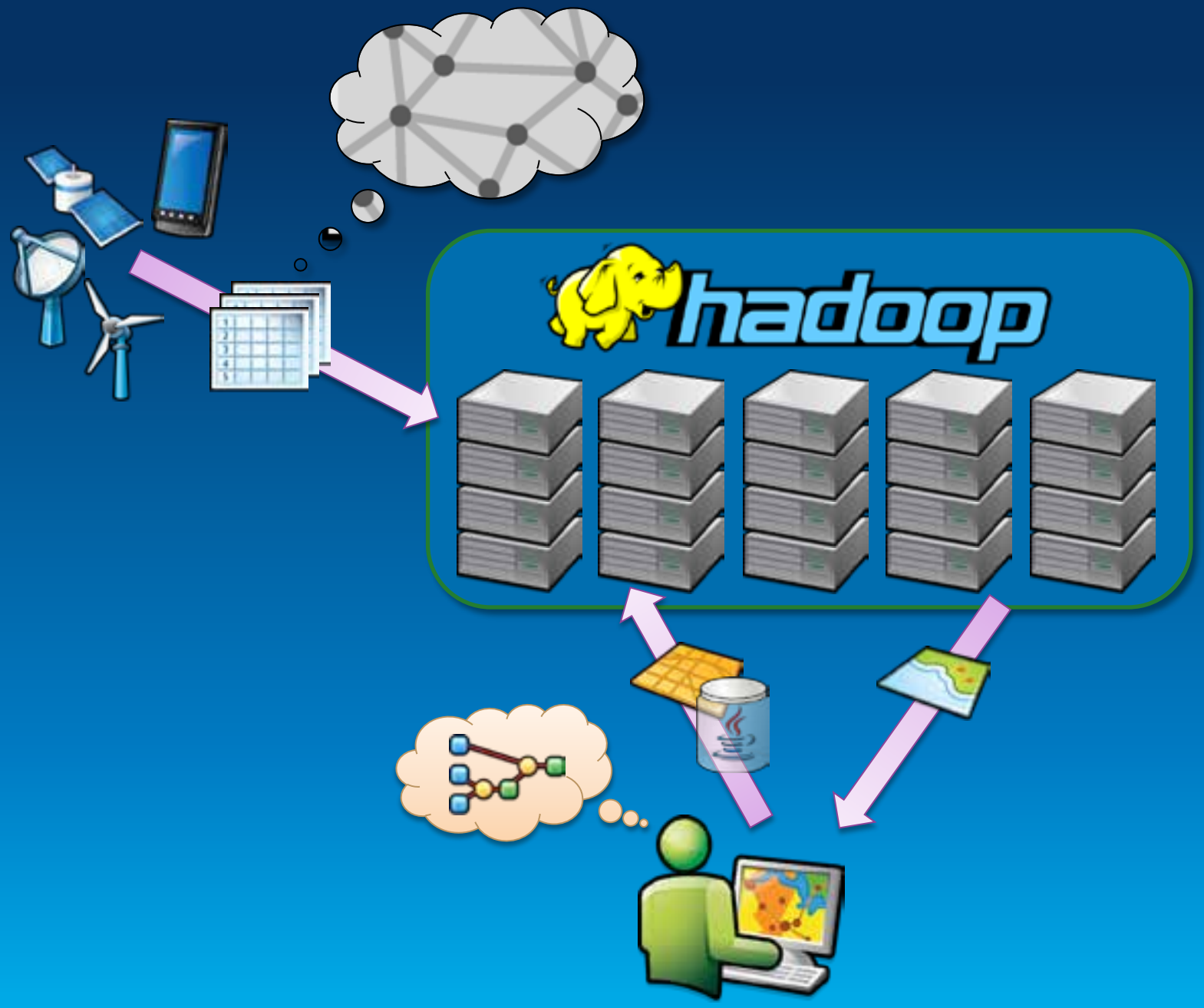
- We didn’t bring 10,000 servers to the user conference…
- … but we do have a small Hadoop cluster we can use for demos

GIS Tools for Hadoop

Credit: Fuqing Zhang and Yonghui Weng, Pennsylvania State University; Frank Marks, NOAA; Gregory P. Johnson, Romy Schneider, John Cazes, Karl Schulz, Bill Barth, The University of Texas at Austin

GIS Tools for Hadoop

- Hadoop is a data processing system that is designed to store and process large amounts of data
- The most common Hadoop data processing task is to reduce a large amount of data to a smaller, more manageable amount of data
- The GIS Tools for Hadoop provide query functions and API methods that enable Hadoop application developers to perform this data reduction process on spatial data

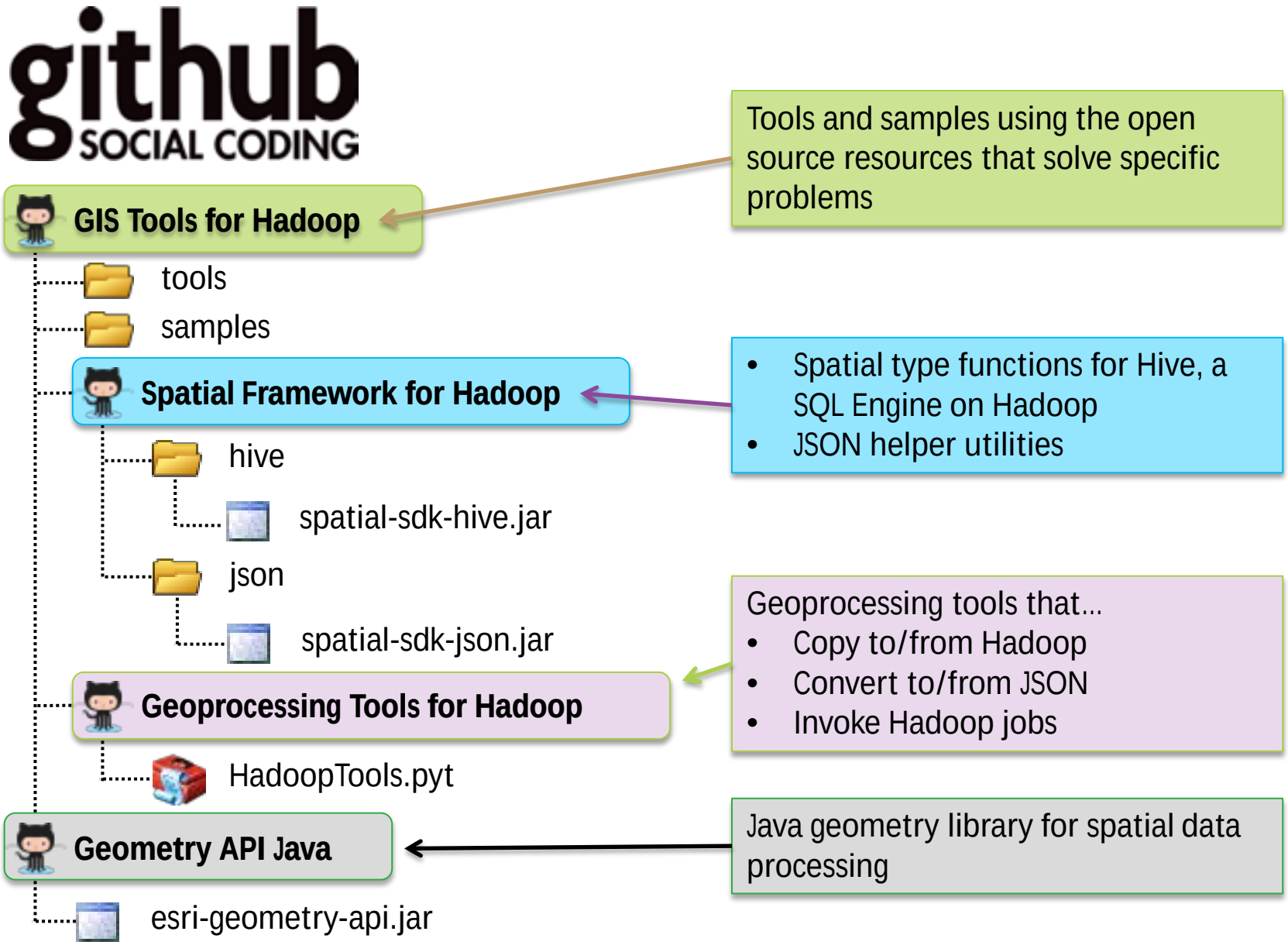

hadoop

Data Reduction Patterns

- Need to reduce large volumes of data into manageable datasets that can be processed in the ArcGIS Platform
  - Filtering
  - Grouping
    - Simple “binning” against grid cells
    - Aggregation into known spatial areas
  - More complex patterns if desired

GIS Tools for Hadoop

- Developer API to support data reduction workflows
  - Release in March 2013 at Esri Developer Summit
- Spatial reduction with full geometry capabilities
  - Relational operators (touch, disjoint, …)
  - Topological operators (buffer, union, intersection, …)
  - Accessible via
    - SQL Expressions
    - Java
  - Small set of GP Tools to migrate data and invoke jobs


github
SOCIAL CODING
GIS Tools for Hadoop
tools
samples
Spatial Framework for Hadoop
hive
spatial-sdk-hive.jar
json
spatial-sdk-json.jar
Geoprocessing Tools for Hadoop
HadoopTools.pyt
Geometry API Java
esri-geometry-api.jar
Tools and samples using the open source resources that solve specific problems
• Spatial type functions for Hive, a SQL Engine on Hadoop
• JSON helper utilities
Geoprocessing tools that…
• Copy to/from Hadoop
• Convert to/from JSON
• Invoke Hadoop jobs
Java geometry library for spatial data processing

DEMONSTRATION

MapReduce Demos

Earthquake Measurement Data

- Data format is CSV, already stored in HDFS file

```
1964/03/28 06:42:57.00,57.98,-151.6,33.0,5.2,ML,0
1964/03/28 06:43:54.40,58.26,-151.25,4.0,6.1,ML,0
1964/03/28 06:50:52.00,56.9,-151.7,33.0,5.1,ML,0
1964/03/28 06:53:35.90,58.79,-149.54,20.0,5.7,ML,0
1964/03/28 07:09:07.60,59.79,-148.1,13.0,5.3,ML,0
1964/03/28 07:10:22.00,58.83,-149.29,17.0,6.2,ML,0
1964/03/28 07:16:16.70,58.0,-150.7,33.0,5.3,ML,0
1964/03/28 07:24:24.60,59.62,-148.7,20.0,5.1,ML,0
```

Coordinates Depth Magnitude

DEMONSTRATION

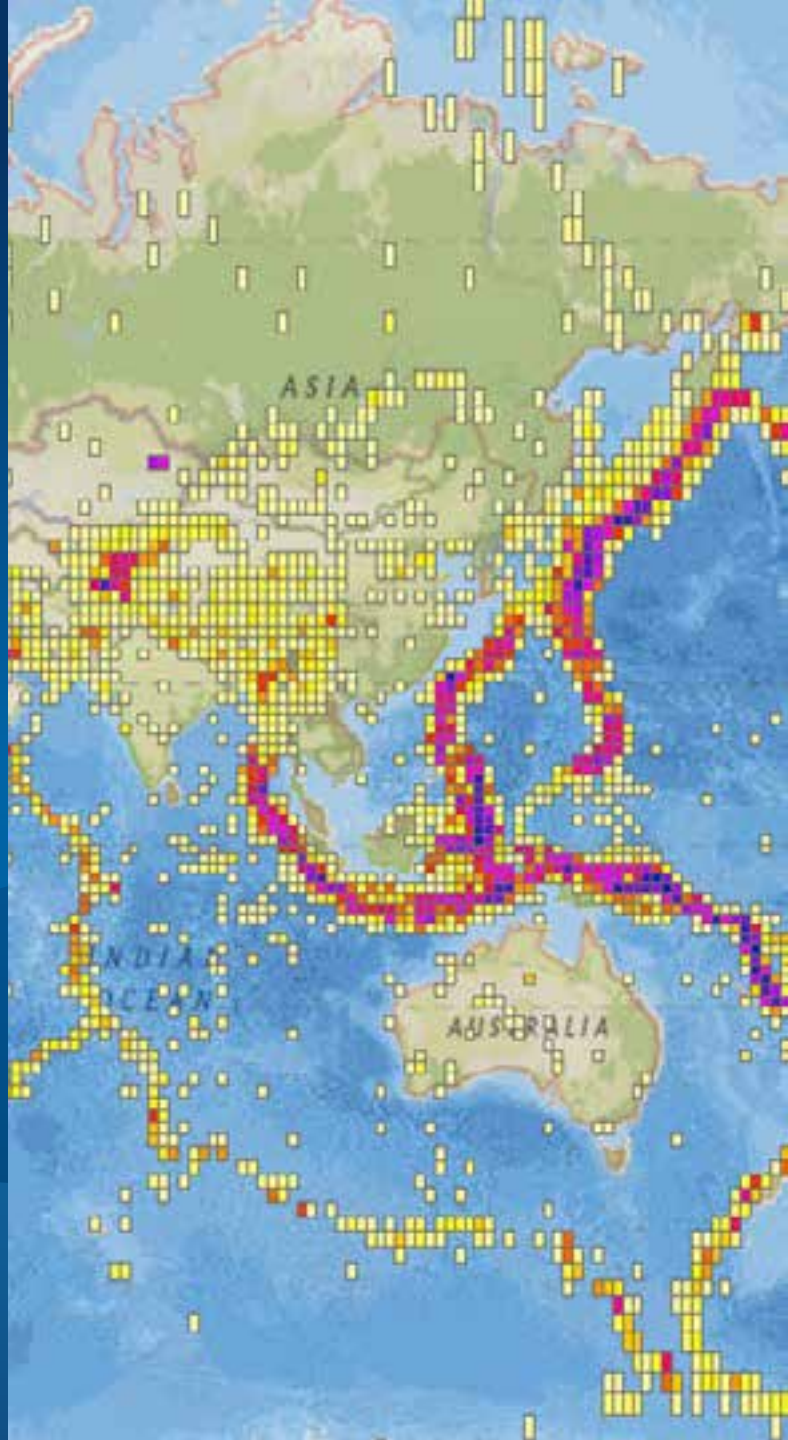
Grouping by Polygons

In this demo, we use ArcGIS to provide the polygonal boundaries of the U.S. States, and the Hadoop application aggregates earthquake data by the state boundaries

DEMONSTRATION

Grouping into Grid Cells

In this demo, we compute arbitrary grid cells such as a 1km grid, and aggregate earthquakes by these cells



The Future

- GIS Tools for Hadoop is a developer story
  - Our initial foray into Big Data
- Esri will continue forward in this domain
  - User stories and technology will be the focus
- Swing by the island to chat further
- Please fill out the evaluations: offering 1330

Questions?

esri

Understanding our world.