Brazilian Federal Police

Brasília University

ForeIndex

Framework for Storage and Indexing of Forensic Data

Marcelo Antonio da Silva
Brazilian Federal Police

Romualdo Pereira
Brazilian Space Agency

Schedule

- Brazilian Federal Police / Brasília University
- Demand for Storage and Indexing in Forensics
- ForeIndex – Workflow
- ForeIndex – Architecture
- Case Study

Brazilian Federal Police

- Forensic Computer Crime Unit – Brasília/DF

DEPARTAMENTO DE POLÍCIA FEDERAL
INSTITUTO NACIONAL DE CRIMINALÍSTICA

Brasília University / Brazilian Federal Police

- Partnership with Universities:
  - **Brasília University**

© Augusto Areal - www.infobrasilia.com.br

Brazilian Federal Police

- Forensic Computer Crime Unit - 2010:
  - Specialists made 9050 reports;
  - Analysis of around 4.6 PB of data on cybercrimes;
  - Some cases with seizures hundreds of computers;
  - Necessity to analyse data correlated of differents medias of the same case.

Demands for Storage and Indexing

- Some cases results in seizure of hundreds of medias;
- Modern foresics tools provide many artifacts for each media analized;
- In some cases this is only data, not knowledge;
- Demand to triage and analisy correlated data.

Demands for Storage and Indexing

- Storage and Indexing as the bottleneck in this process;
- Case Study:
  - 2.274.796 files (482 GB);
  - OS: Windows 7 / openSUSE 11.4 (Linux 2.6)
  - Hardware: Intel Core-2 Quad, 2.66 GHz, 4 GB RAM
  - Average Time to Copy: **12 hours** (NTFS, Ext4)
  - Average Time to Index: **26 hours** (Forensics Tools)

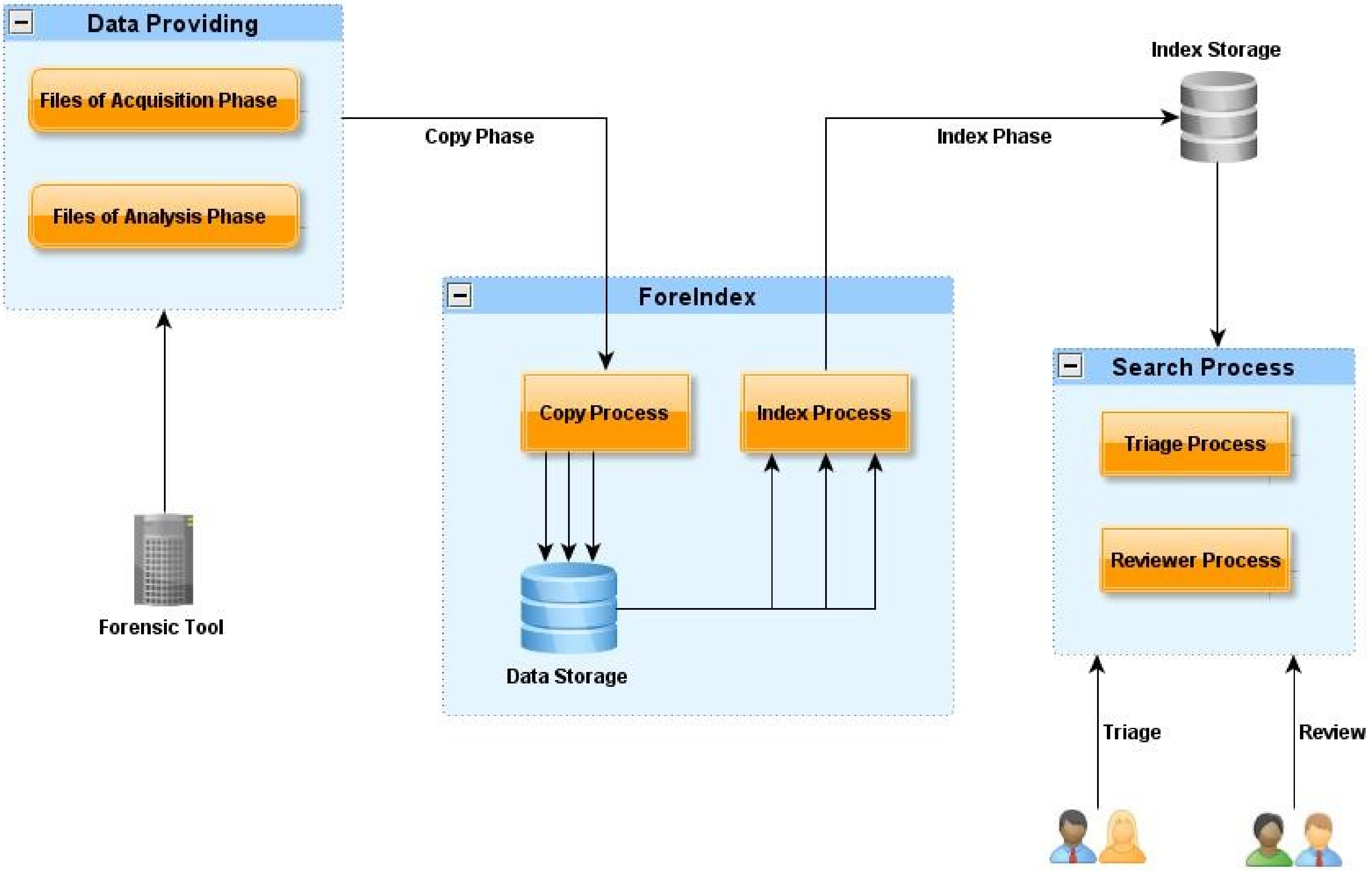
ForeIndex

- Framework for storage and indexing distributed of Forensic Data;

- Utilized in 2 cases of the forensics process:
  - After of the data acquisition phase (triage):
    - Minimize the amount to data to analyse.

  - After of the analysis phase (reviewers):
    - Enabling analysis of correlated evidence.

ForeIndex - WorkFlow



ForeIndex – WorkFlow (Tools)

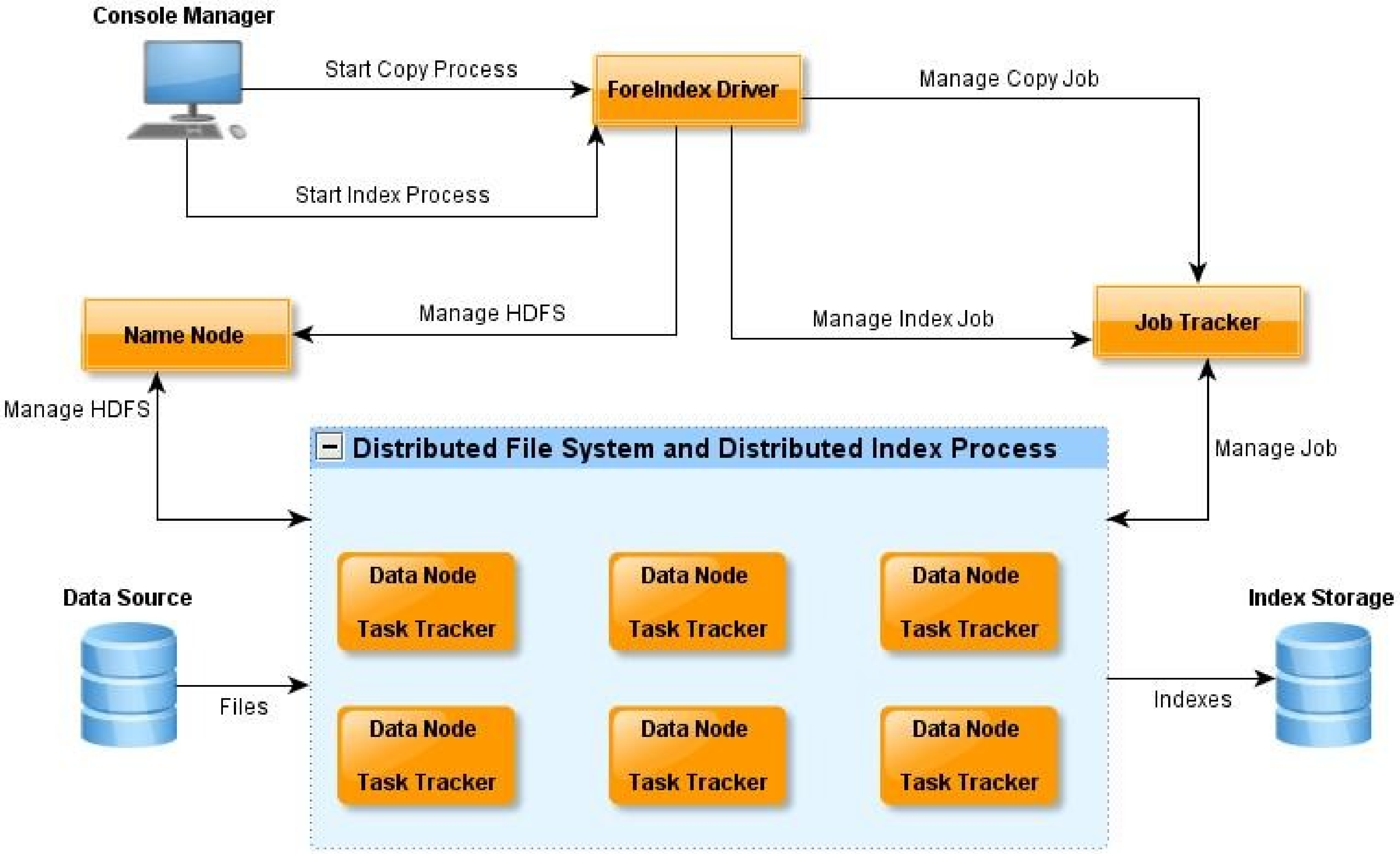
- Data Providing:
  - Sleuth Kit Scripts;
  - Files of another forencisc analysis process.

- Search Process:
  - Apache Solr;
  - JSP and Servlets (Smart GWT);
  - Jasper Reports.

ForeIndex – WorkFlow (Tools)

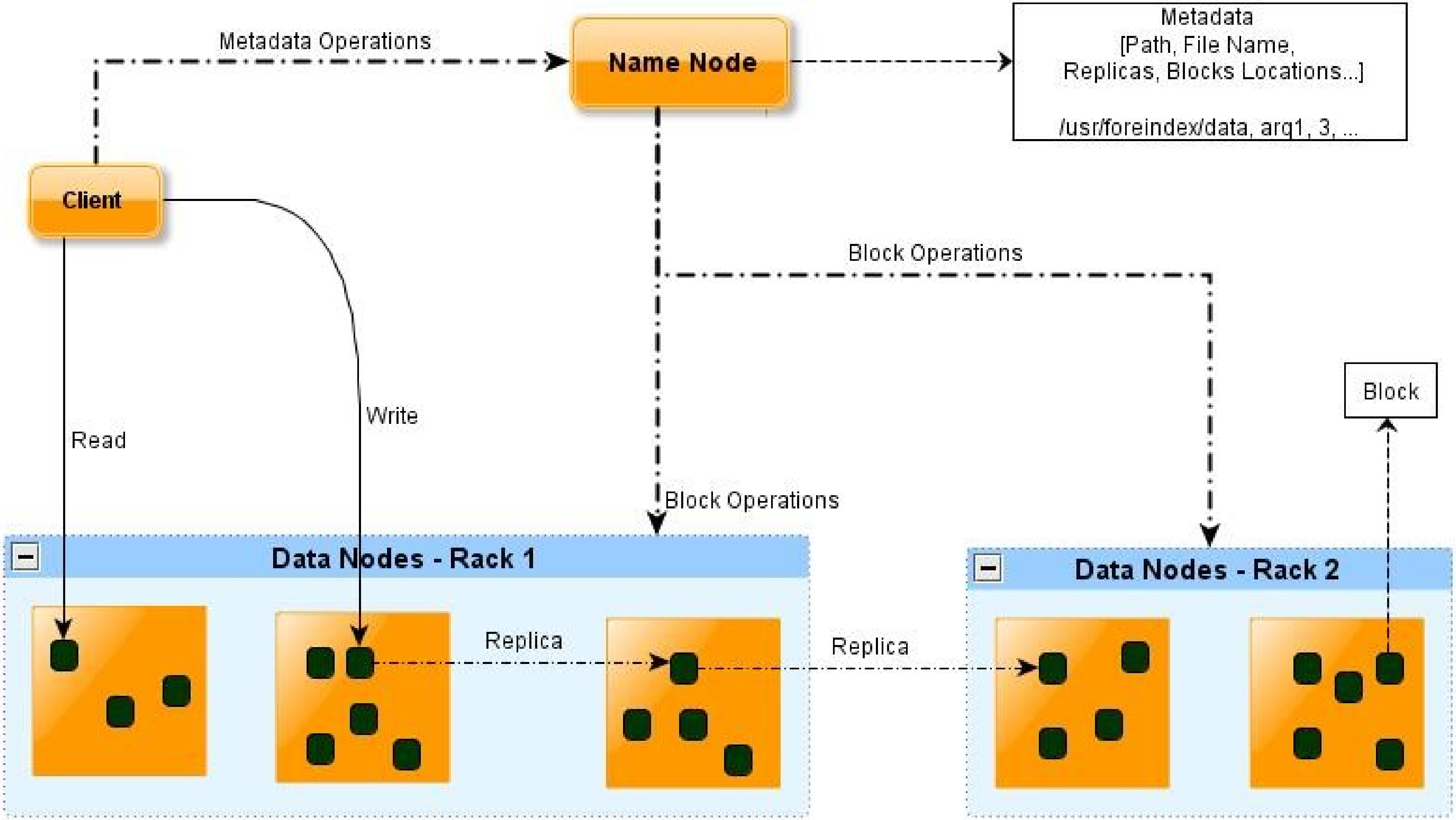
- ForeIndex:
  - Copy and Index Phases;
  - Hadoop Distributed FileSystem (HDFS)
  - Hadoop MapReduce;
  - Lucene Indexer;
  - Tika.

ForeIndex - Architecture



ForeIndex – Architecture

HDFS Architecture



ForeIndex – Architecture

HDFS Architecture

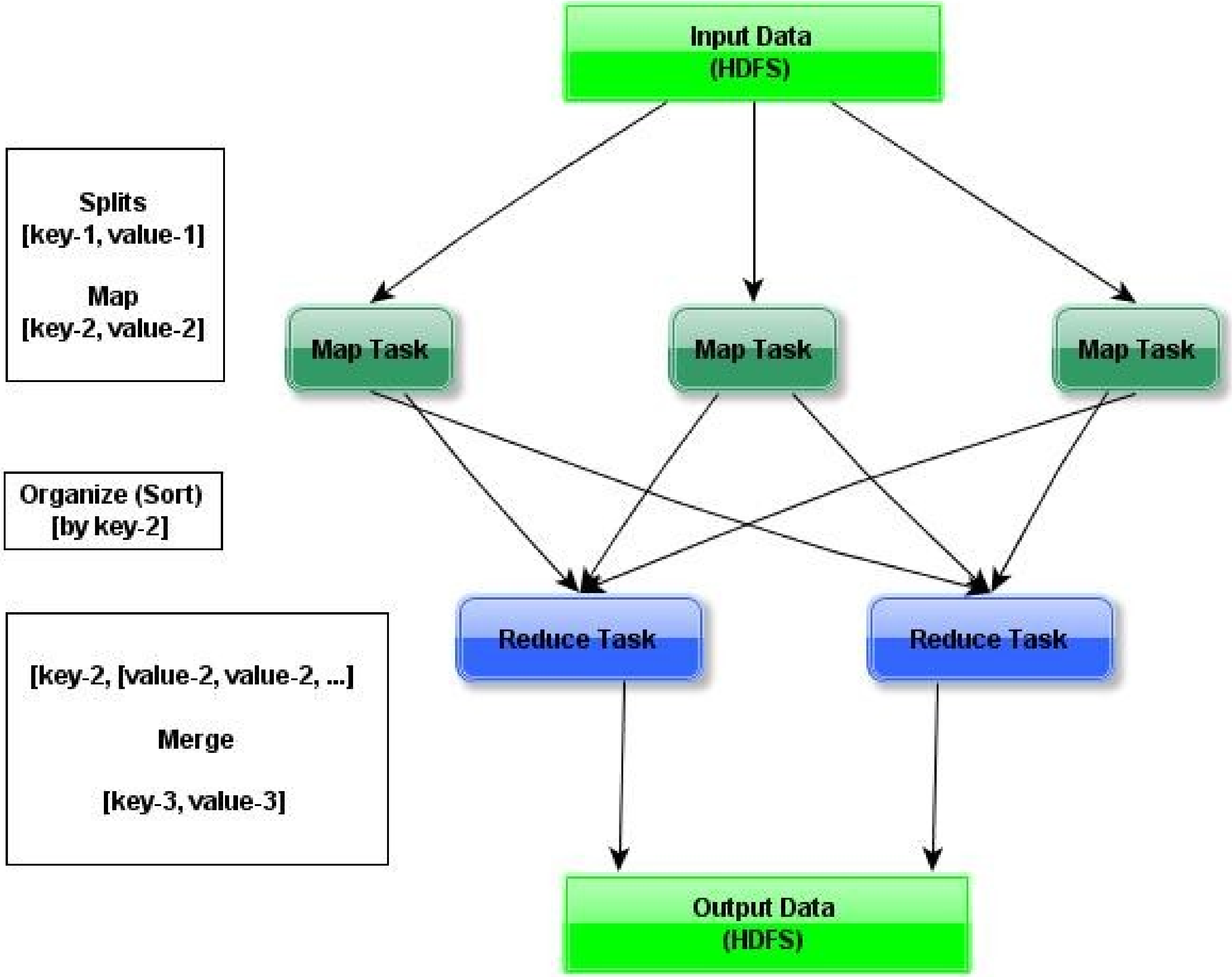
- HDFS – Features:
  - Streaming data access;
  - Commodity hardware;
  - Namenode and Datanodes;
  - Data Replication;
  - Data Blocks;
  - Data disk failure, heartbeats, re-replication

ForeIndex – Architecture
MapReduce

- MapReduce - Google™:
  - Parallel programming model for data processing;
  - Processing in 2 phases: Map Phase, Reduce Phase;
  - Commodity hardware, fault-tolerant manner;
  - Data input splitted for map tasks processing;
  - Maps output organized and processed for reduce tasks;
  - Maps and Reduces tasks are schedulled and monitored;
  - Compute nodes and datanodes tipically are the same

ForeIndex – Architecture

MapReduce



Splits
[key-1, value-1]

Map
[key-2, value-2]

Organize (Sort)
[by key-2]

[key-2, [value-2, value-2, ...]

Merge

[key-3, value-3]

Input Data
(HDFS)

Map Task

Map Task

Map Task

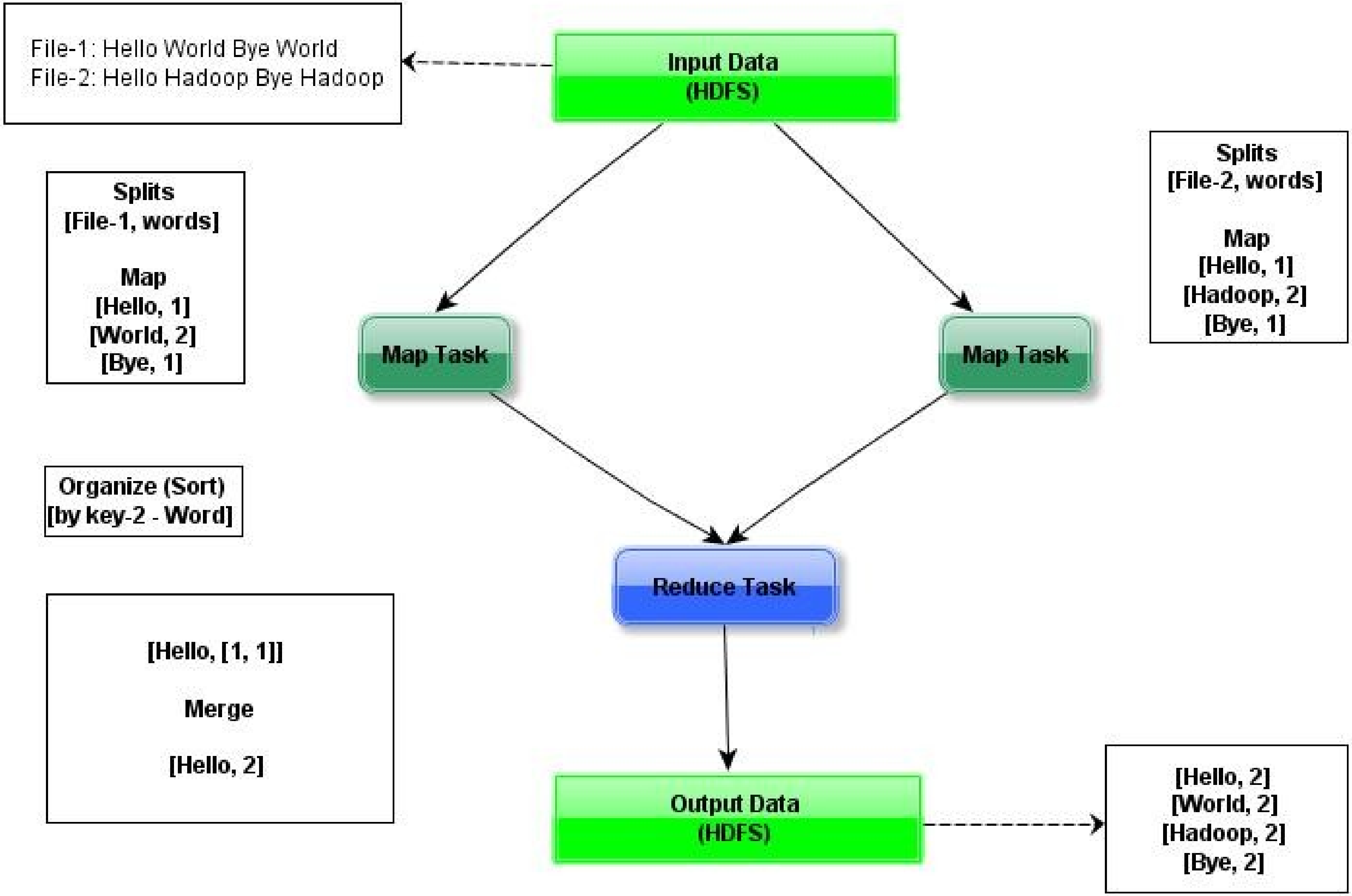
Reduce Task

Reduce Task

Output Data
(HDFS)

ForeIndex – Architecture

MapReduce – Example (WordCount)



File-1: Hello World Bye World
File-2: Hello Hadoop Bye Hadoop

Input Data
(HDFS)

Splits
[File-1, words]

Map
[Hello, 1]
[World, 2]
[Bye, 1]

Map Task

Map Task

Splits
[File-2, words]

Map
[Hello, 1]
[Hadoop, 2]
[Bye, 1]

Organize (Sort)
[by key-2 - Word]

Reduce Task

[Hello, [1, 1]]

Merge

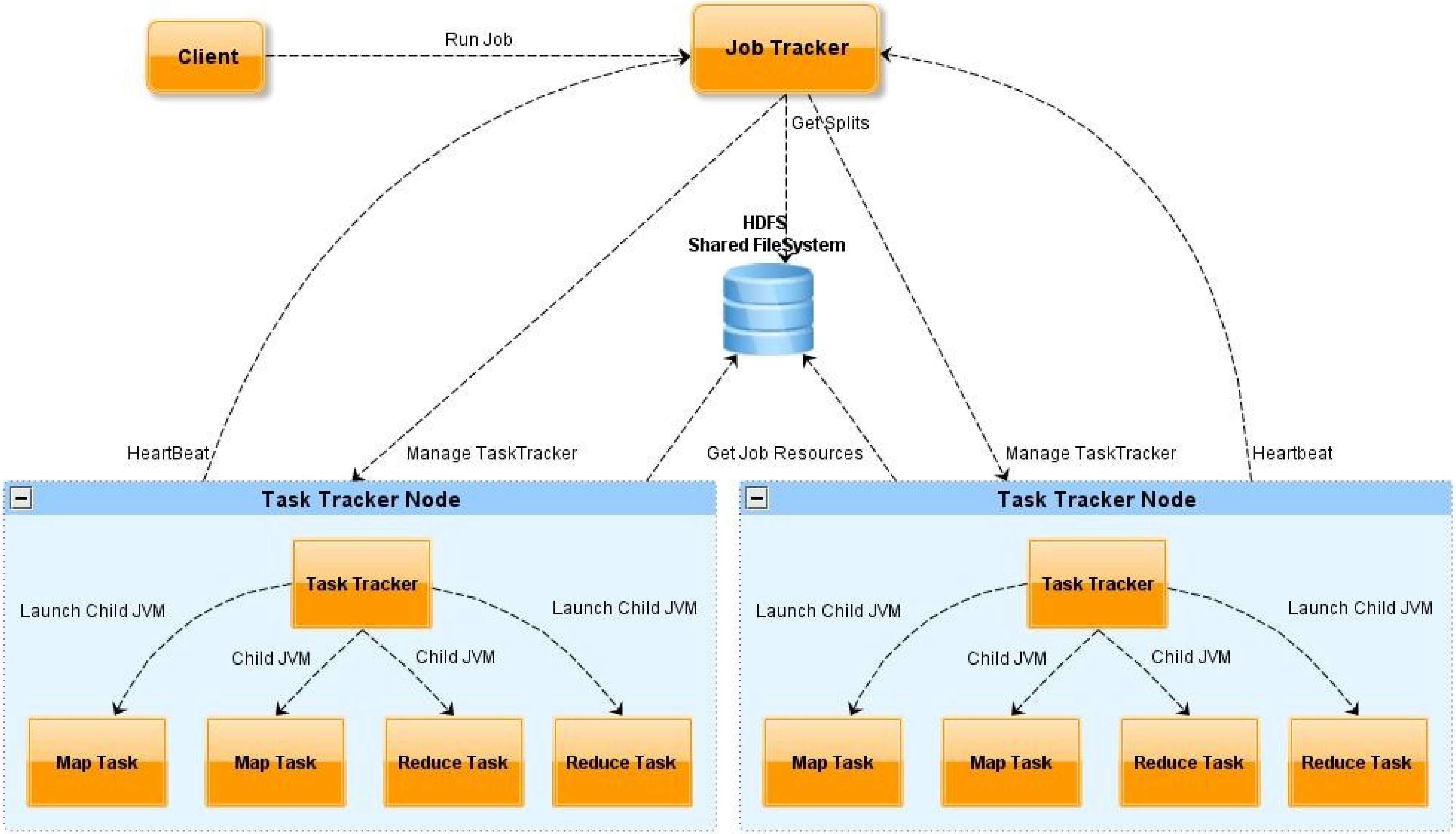
[Hello, 2]

Output Data
(HDFS)

[Hello, 2]
[World, 2]
[Hadoop, 2]
[Bye, 2]

ForeIndex – Architecture

MapReduce – Hadoop Architecture

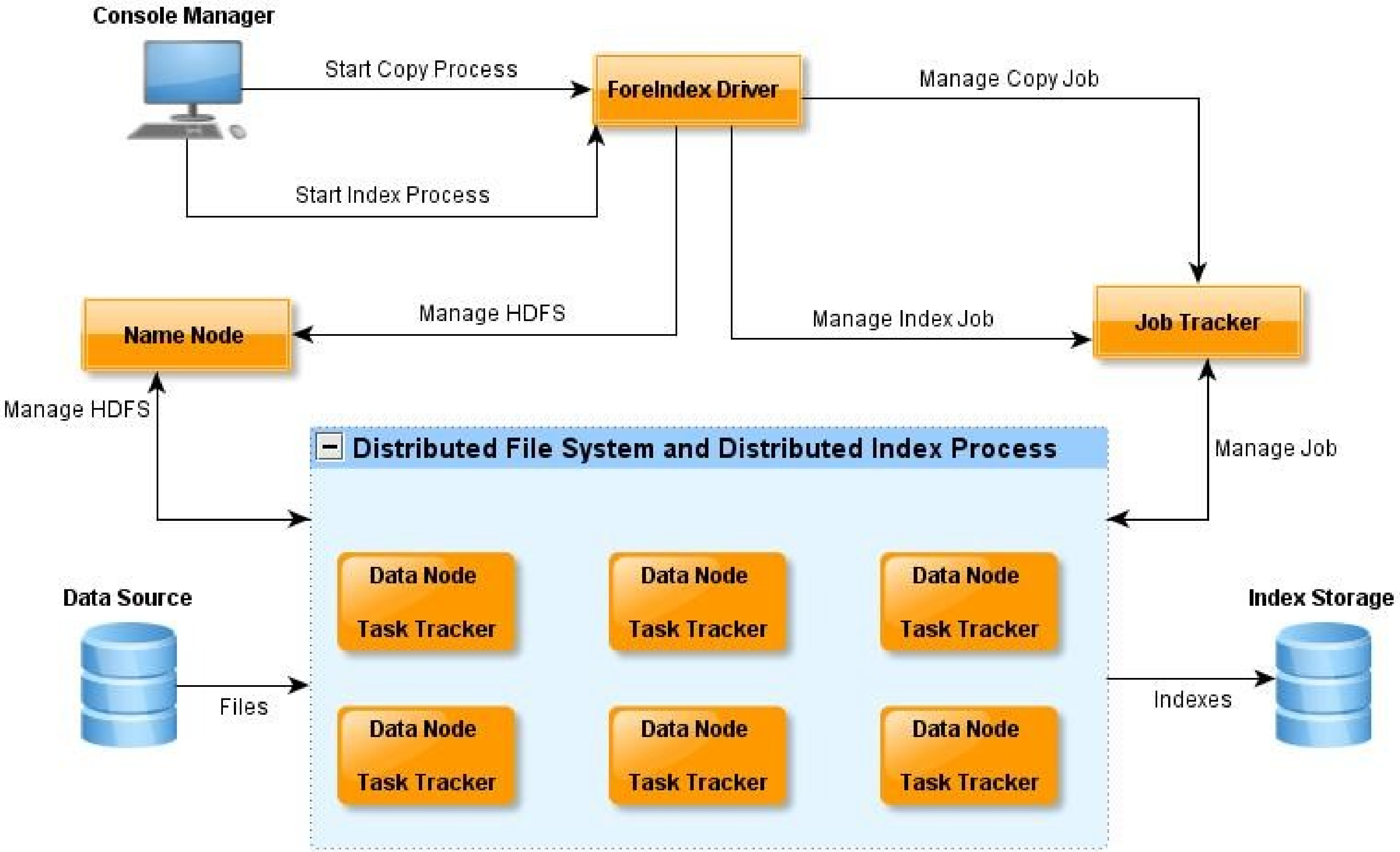


ForeIndex – Architecture

MapReduce – Hadoop

- Hadoop MapReduce – Features:
  - HDFS Block Size – Input Split Size;
  - Data Locality;
  - Job Manage and Monitoring;
  - MapReduce functions in many languages;
  - Many data types and formats;
  - Counters, Sorter, Joins.

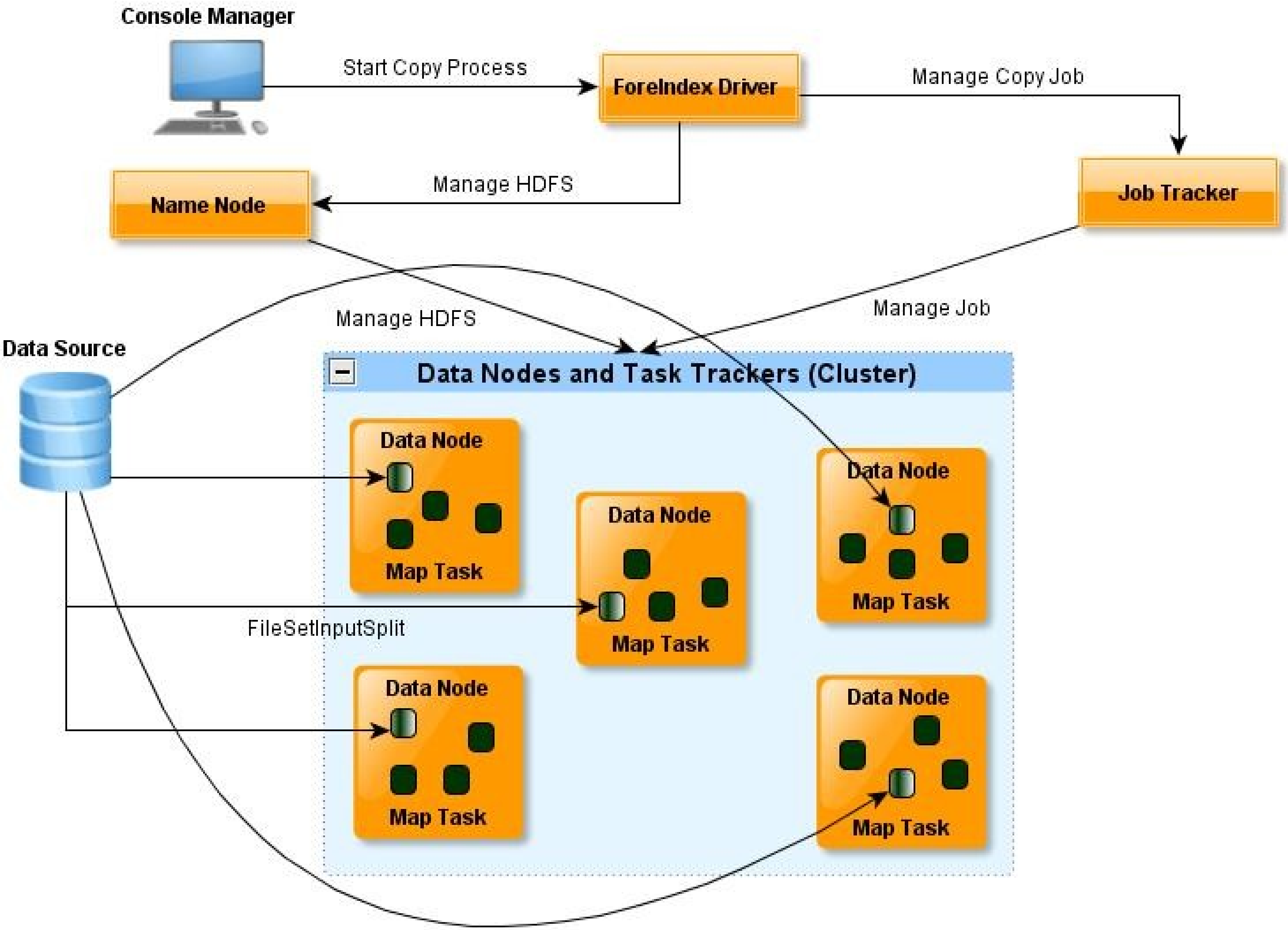
ForeIndex - Architecture



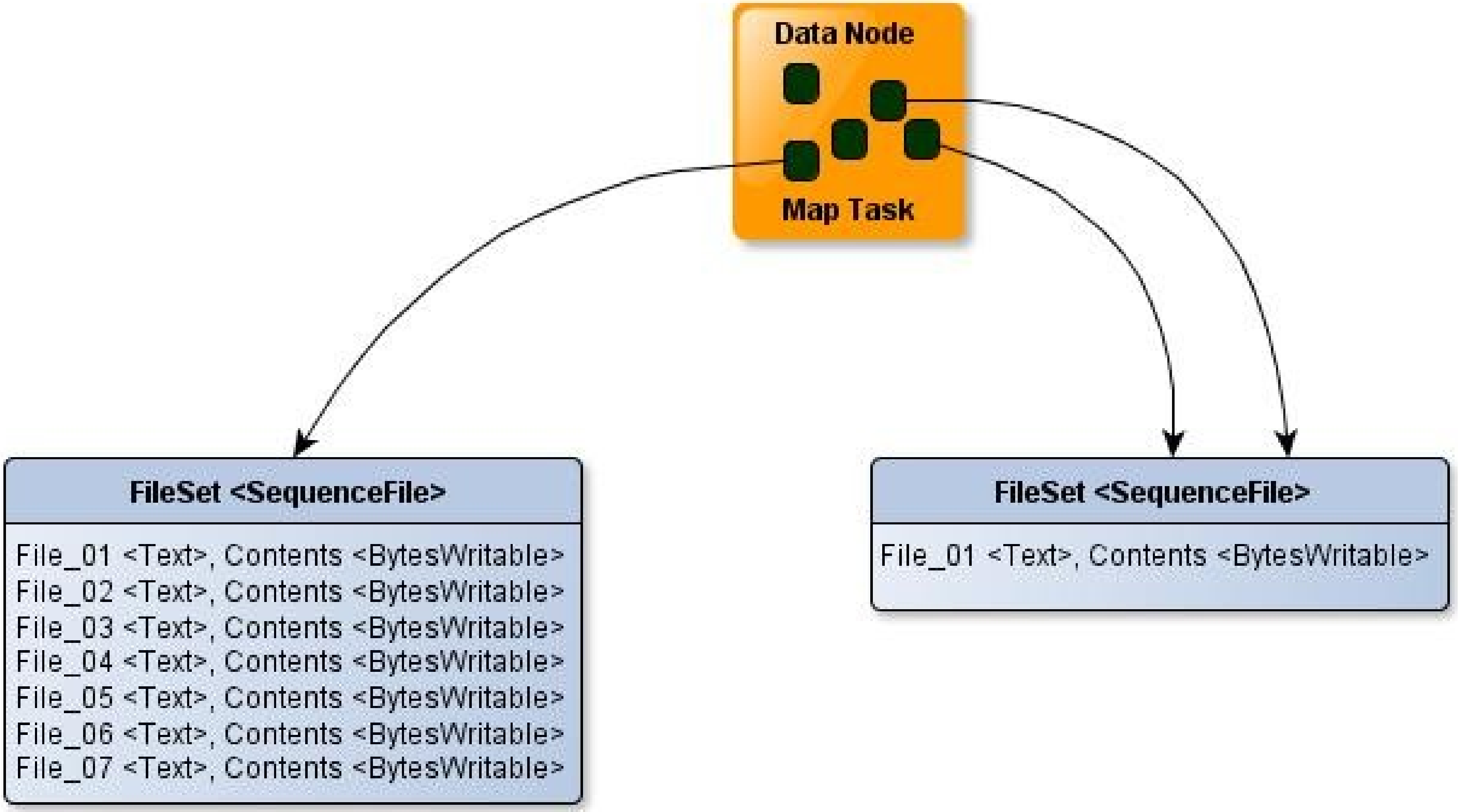
ForeIndex – Copy Process

- Copy Process - Requirements:
  - Many files to process;
  - Many types os files to process;
  - File size less than block size (in average);
  - Block Size in HDFS is similar of Cluster Size in NTFS;
  - Majority of files can’t be splitted to be parsed.

ForeIndex – Copy Process



ForeIndex – Copy Process



Data Node

Map Task

FileSet <SequenceFile>

File_01 <Text>, Contents <BytesWritable>
File_02 <Text>, Contents <BytesWritable>
File_03 <Text>, Contents <BytesWritable>
File_04 <Text>, Contents <BytesWritable>
File_05 <Text>, Contents <BytesWritable>
File_06 <Text>, Contents <BytesWritable>
File_07 <Text>, Contents <BytesWritable>

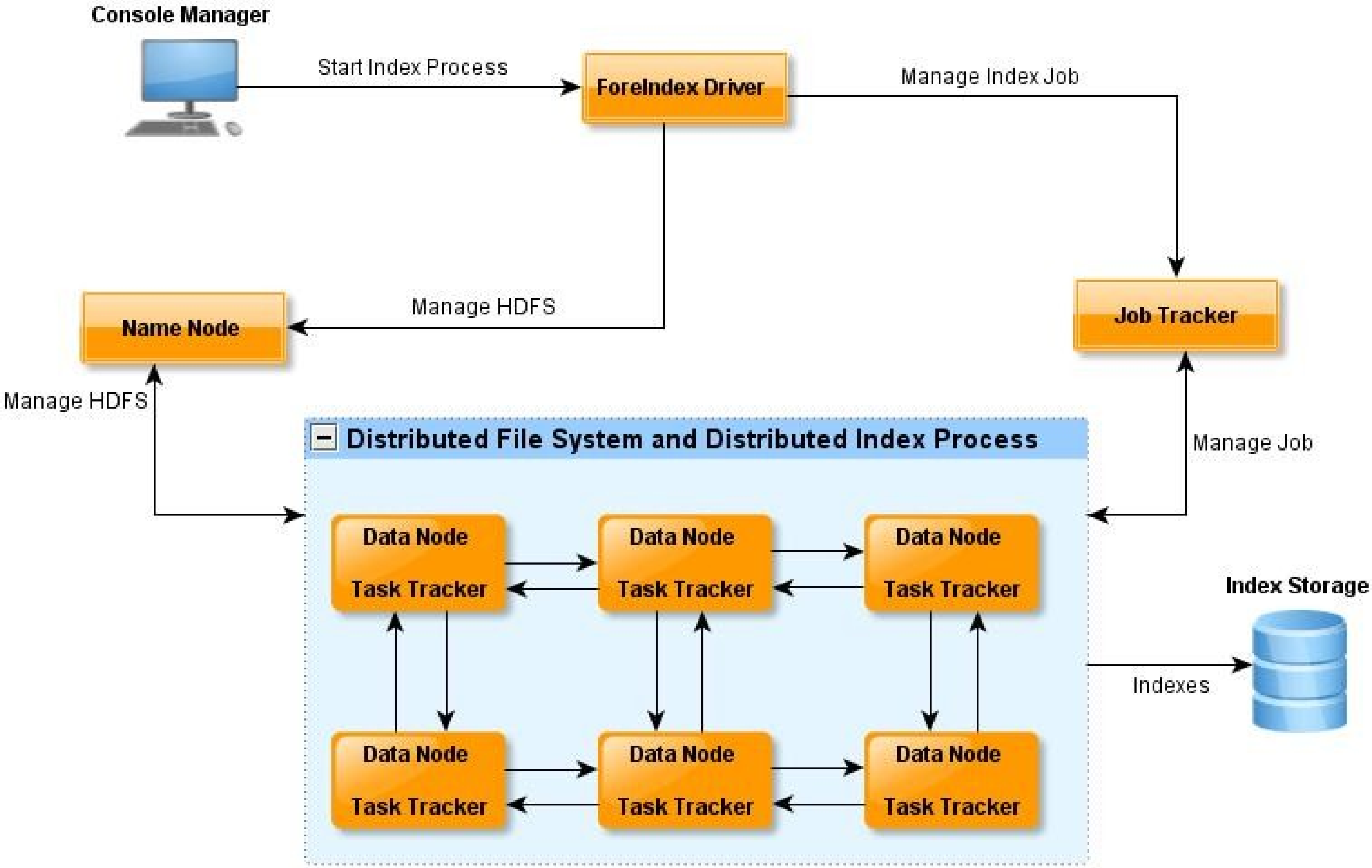
FileSet <SequenceFile>

File_01 <Text>, Contents <BytesWritable>

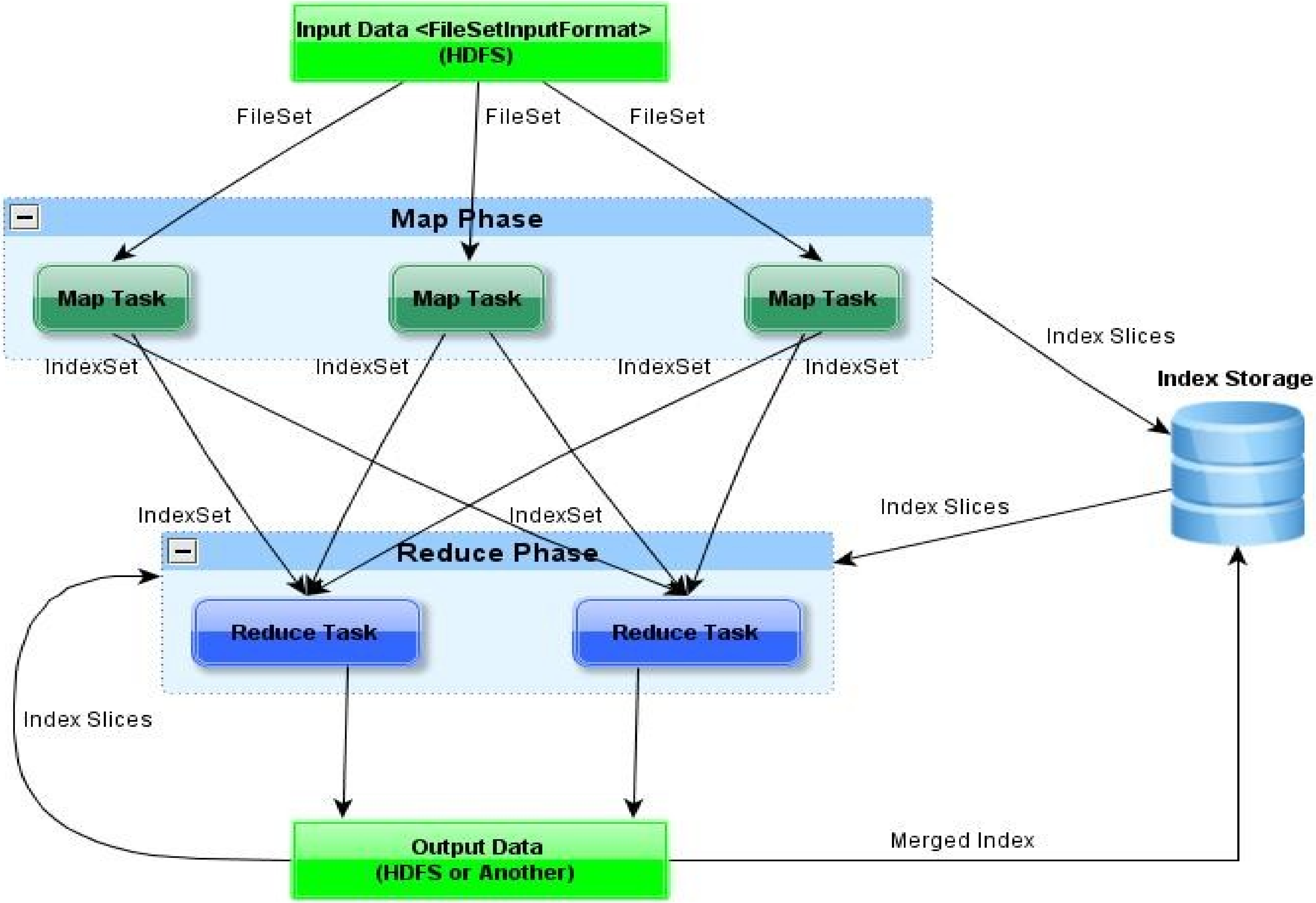
ForeIndex – Copy Process

- Features:
  - Distributed copy process;
  - One or more files contained in FileSet <SequenceFile>;
  - FileSet at least with the size of HDFS Block;
  - Namenode more efficientily used;
  - File is not splitted (good for parsing);
  - Other benefits in indexing process.

ForeIndex – Index Process



ForeIndex – Index Process



ForeIndex – Index Process

- Features:
  - Distributed index process;
  - Use Lucene and Tika;
  - Input data are SequenceFiles (FileSet);
  - Sequence Files and Data Locality;
  - Pipeline for read, parse and index the files;
  - The index slices are a functional index;
  - The index slices can be merged.

ForeIndex – Case Study

- Standalone Test:
  - 2.274.796 files (482 GB);
  - OS: Windows 7 / openSUSE 11.4 (Linux 2.6)
  - Hardware: Intel Core-2 Quad, 2.66 GHz, 4 GB RAM
  - Average Time to Copy: **12 hours** (NTFS, Ext4)
  - Average Time to Index: **26 hours** (Forensics Tools)
  - Time to Copy in Forensic Cloner: 02:40 (hh:mm)

ForeIndex – Case Study

- ForeIndex Test (2.274.796 files)
  - Configuration:
    - 2.274.796 files (482 GB);
    - Files format: .txt, .xls(s), .xls, .doc(x), .rtf, .msg
    - OS: openSUSE 11.4 (Linux 2.6)
    - Hardware: Intel Core-2 Quad, 2.66 GHz, 4 GB RAM
    - Cluster: 12 Machines (1 Namenode, 1 Job Tracker, 10 Workers [Datanode, TaskTracker]);
    - HDFS Block Size: 64 MB;
    - Local Area Network: 1 Gbps;

ForeIndex – Case Study

- ForeIndex Test (2.274.796 files)
  - Copy Process:
    - Data Source: 2 HDDs SATA-II (no RAID);
    - 4 Maps per Worker = 40 Maps;
    - SequenceFiles created = 206.799;
    - Time to copy = **03:25** (hh:mm)
    - Time to Copy in Forensic Cloner: **02:40** (hh:mm)
    - Time to Copy in Standalone Test: **12:00** (hh:mm)

ForeIndex – Case Study

- ForeIndex Test (2.274.796 files)
  - Copy Process:
    - Data Source: 2 HDDs SATA-II (RAID-1);
    - 4 Maps per Worker = 40 Maps;
    - SequenceFiles created = 206.799;
    - Time to copy = **01:50** (hh:mm)
    - Time to Copy in Forensic Cloner: **02:40** (hh:mm)
    - Time to Copy in Standalone Test: **12:00** (hh:mm)

ForeIndex – Case Study

- ForeIndex Test (2.274.796 files)
  - Index Process:
    - 30 Maps, 10 Reducers;
    - SequenceFiles processed = 206.799;
    - Time to Index in Standalone Test = **26:00** (hh:mm)
    - Time to Index in ForeIndex Cluster = **00:25** (hh:mm)

Questions?

DEPARTAMENTO DE POLÍCIA FEDERAL
INSTITUTO NACIONAL DE CRIMINALÍSTICA

Marcelo Antonio da Silva

Brazilian Federal Police

marcelosilva.mas@dpf.gov.br

Brazilian Federal Police

Brasília University

ForeIndex

Framework for Storage and Indexing of Forensic Data

Marcelo Antonio da Silva
Brazilian Federal Police

Romualdo Pereira
Brazilian Space Agency