Each week I’ll ask you to pick a KeyWord, Phrase, Term, etc and to add a few sentences to it. This will be a class exercise before you comme

nce the Lab Exercises.

You will do this each week.

Over the course of the semester the class will have created a valuable resource as a take-away from the module. It will help with remembering these terms and may be of help when preparing for the end of Semester MCQ.

Dictionary Week 9

**Complete the work listed below for you BreakOut Room**

| **BreakOut**  **Room #** | **Word/Phrase/Term, etc** | **Description explaining the meaning** | **Students who Contributed to this** |
| --- | --- | --- | --- |
| 1 | Spark Core | Contains the basic functionality of Spark: - components for task scheduling,  - memory management to deliver speed of execution,  - fault recovery,  - managing storage systems,  - a support model for applications  - provides Java, Scala, and Python APIs to aid the development process  - APIs that also define RDDs (Spark’s main programming abstraction) | Ciaran Finnegan |
| 2 | Spark Session | A unified entry point for all Spark processing which simplifies setup and reduces complexity. It facilitates working with RDD, DataFrame and Dataset and includes the APIs in different contexts (SparkContext, SQLContext, StreamingContext and HiveContext). | Niamh Clarke, Trevor Doherty |
| 3 | Spark Job (DAG) | A Directed Acyclic Graph (or DAG) is a representation of the logical execution of a Spark job, consisting of stages and tasks. The DAG shows the dependencies between stages and tasks, indicating which tasks must be completed before others can begin. | Simon Hall |
| 4 | RDD | An RDD is a fundamental data structure that represents an immutable collection of elements that can be processed. RDDs are the primary building blocks of Spark applications partitioned across nodes in your cluster that can be operated in parallel with a low-level API that offers transformations and actions. | Nishant Kapur, Marc O Neill |
| 5 | Lazy Evaluation | Operations are delayed until their value is needed, as opposed to strict evaluation where they’re run after each command. In Spark, this means that transformations can be run but a process will only be executed when an action is called. | Niall Clarke |
| Any other words you can add? |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Dictionary for Week 4**

**Complete the work listed below for you BreakOut Room**

| **BreakOut**  **Room #** | **Word/Phrase/Term, etc** | **Description explaining the meaning** | **Students who Contributed to this** |
| --- | --- | --- | --- |
|  |  |  |  |
| 1 | Pig Laten | The Pig Latin is a data flow language used by Apache Pig to analyze the data in Hadoop. | NishantK. |
| 2 | Hive | Hive is a data warehouse system that enables analytics at a large scale. It is built on top of Hadoop. It provides a simple interface for ad hoc querying, analyzing and summarizing large  amounts of data using a SQL like language.  The target audience for Hive is data analysts and people who already have SQL expertise. | Rachel D, Niamh |
| 3 | Hive Bucket | Bucketing in hive is the concept of breaking data down into ranges, which are known as buckets, to give extra structure to the data so it may be used for more efficient queries. The range for a bucket is determined by the hash value of one or more columns in the dataset |  |
| 4 | HBase Column Family | Hbase is a column-oriented, non-relational database management system that runs on top of Hadoop Distributed File System. HDFS.HBase provides a fault-tolerant way of storing sparse data sets, which are common in many big data use cases.  An HBase table is made of column families which are **t**he logical and physical grouping of columns. The columns in one family are stored separately from the columns in another family. If you have data that is not often queried, assign that data to a separate column family. | Joel Quadras, Ciaran Finnegan |
| 5 | Sqoop | Sqoop (SQL to Hadoop) is a tool for bulk transferring data between databases within Hadoop (using MapReduce at an underlying level).Incremental loads of a single table or free form SQL queries are supported, as well as saved jobs which can be run multiple times to import updates to the database since the last import. | Niall Clarke, Trevor Doherty |
| 6 | Flink | Apache Flink is a stream-processing and batch-processing framework typically used in conjunction with a wide variety of input output systems such as HDFS or Kafka and comes bundled with Hadoop MapReduce. Flink has been designed to be used with continuously produced data. | Marc O’Neill, Simon Hall |
| Any other words you can add? | Mahout |  |  |
|  |  |  |  |
|  |  |  |  |

**Dictionary for Week 3**

**Complete the work listed below for you BreakOut Room**

| **BreakOut**  **Room #** | **Word/Phrase/Term, etc** | **Description explaining the meaning** | **Students who Contributed to this** |
| --- | --- | --- | --- |
| 1 | Yarn | YARN (Yet Another Resource Negotiator) is the resource management layer of Hadoop that allocates resources like CPU, memory and storage to applications running on a Hadoop cluster. YARN provides a platform-agnostic way to run multiple applications on a Hadoop cluster, allowing for flexibility in the types of big data workloads that can be executed. This enables better utilization of cluster resources and improves the scalability and performance of Hadoop. | JQ, NK, CF |
| 2 | Combiners | Due to Mappers potentially generating large data volumes, this can result in significant network traffic when passed to Reducers. As such, Combiners can help alleviate this issue by acting as a mini-Reducer. They can reduce the intermediate data being passed to Reducers consequently reducing its workload and leading to faster jobs and decreased network traffic. Combiners are run locally on a single Mapper’s output and often use the same code as the Reducer. | TD |
| 3 | Partitionars | Practitioners takes major part in disseminating the key value pairs from the map stage to the reduce stage using hash function (default). This derives the partitions letting it work with the outputs of the mapper. | VM, SNN |
| 4 | Chaining | Chaining is the process of combining multiple mapreduce jobs into a single workflow. This can be useful in reducing complex jobs to smaller tasks. The use of chaining can allow for better scaling with more parallel processing. | Marc, Niall |
| 5 | JobControl | [JobControl](https://hadoop.apache.org/docs/r1.2.1/api/org/apache/hadoop/mapred/jobcontrol/package-summary.html) is a utility which encapsulates a set of MapReduce jobs and their dependencies. | RM |
| 6 | Role of Counters | Counters track the progress of a MapReduce job.  There are built in counters and users can define their own within both mappers and reducers. Often used for troubleshooting or tracking the statistics of program executions. | RD  AM |
| Any other words you can add? | Chained Job Dependencies |  |  |
|  | ToolRunner |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Dictionary for Week 2**

**Complete the work listed below for you BreakOut Room**

| **BreakOut**  **Room #** | **Word/Phrase/Term, etc** | **Description explaining the meaning** | **Students who Contributed to this** |
| --- | --- | --- | --- |
| 1 | MapReduce | Mapreduce is a programming model used to process large datasets in parallel using distribution.With MapReduce, rather than sending data to where the application or logic resides, the logic is executed on the server where the data already resides, to expedite processing. | VM, JQ, NK |
| 2 | Mapper | The Mapper is a function which processes the input data by creating several small chunks of data. The input to the Mapper is in the form of a key/value pair. It produces >=0 outputs (key:value pairs) for each input pair. For example, a counting function may generate multiple key/value pairs - one per element that is counted. | TD |
| 3 | Reducer | The reducer processes intermediate data. It uses the data from the mapper as its input and runs a function over them. Each reducer runs in parallel and the user can specify how many. After processing the reducer will using the map function will produce the output | AM |
| 4 | Shuffle & Sort | Shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce. Sort phase in MapReduce covers the merging and sorting of map outputs. Data from the mapper are grouped by the key, split among reducers, and sorted by the key. Every reducer obtains all values associated with the same key. As shuffling can start even before the map phase has finished so this saves some time and completes the tasks in lesser time. | RM & CB |
| 5 | Job Tracker | The MapReduce JobTracker initialises jobs across multiple nodes in a Hadoop cluster, receives updates (heartbeats) from the TaskTracker, and reassigns in the case of failure. It is a point of failure for MapReduce. | Niall Clarke, Simon Hall |
| 6 | Task Tracker | A TaskTracker is a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations - from a JobTracker.  Every TaskTracker is configured with a set of slots, these indicate the number of tasks that it can accept. | Srishruthi Nandigam Narayanan, Rachel Donnelly |
| Any other words you can add? | YARN |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Week 1**

Pick a key Word, Phrase, Term, etc from this week and write a couple of sentences about it in the table below.

I’ve pre-populated the table with some words. You can add some more

| ***Word/Phrase/Term, etc*** | ***Write a short description explaining its meaning*** | ***Students who contributed to this*** |
| --- | --- | --- |
| Node | A node is a process (i.e. a computer processing or storing) on a VM or machine. A node is a device or a data point in a larger network. Each node will have storage device(s) attached to it. There are two types of nodes in Hadoop - a name node and a data node. | AM, TD,JQ |
| Hadoop | Hadoop is open-source software designed for scalable, distributed computing.  Framework designed to handle the distributed processing of massive amounts of data using simple programming models. | Simon Hall  NK |
| Hadoop the Toy | An elephant doll possessed by Doug Cutting’s son. Doug Cutting is the creator of Hadoop. The name of the doll was Hadoop and the technology was named after it. Cutting wanted a name that was easy to spell, pronounce and unique. | Vyoma Mohan |
| HDFS | The Hadoop Distributed File System (HDFS) is the primary data storage system used by [Hadoop](https://www.techtarget.com/searchdatamanagement/definition/Hadoop) applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters. | RM, SNN |
| Name Node | Management node of a HDFS cluster. Keeps track of metadata - which blocks make up a file. Also knows which datanodes store which blocks. | Marc O’Neill |
| Data Node | A data node is a slave node which stores the data. When it gets instruction from name node or client, it then delivers the data. It needs to communicate with name node periodically to inform it of the data it has stored. | Abhishek Mandloi |
| Cluster | A collection of computers/nodes. Hadoop clusters are composed of a network of master and worker nodes that execute various jobs across the HDFS. Hadoop clusters can boost the processing speed of many big data analytics jobs | Sean Mc |
|  |  |  |
|  |  |  |