GLOSSARY

Big Data Analytics with Hadoop and Apache Spark

With Kumaran Ponnambalam

Use these terms and definitions below to understand concepts taught in the course.

Transcript Search: note that you can search for terms directly within the course. To search video text, switch to the *Transcripts* tab, then press Cmd/Ctrl + F on your keyboard to run a search within the active transcript.

Term	Definition
Apache Spark	An open-source technology that provides a large-scale distributed data processing engine in which data is primarily stored in memory
bucketing	A partitioning technique that is used when there are a large number of unique values for a given column, and can improve performance in certain data transformations by avoiding data shuffling and sorting
compression	A process of encoding information using fewer bits than the original representation, with a goal to save storage capacity, speed up file transfer, and decrease costs for storage hardware and network bandwidth
HDFS	Hadoop Distributed File System; an open-source technology that provides distributed data storage and computing using low-cost software
joins	Combining two or more different tables (sets) to get one set of the result based on some criteria; Spark offers most of the commonly used joins in SQL
partitioning	A technique that divides large amounts of data into multiple slices based on specific column values, and provides a way to read only a subset of data; optimal when a given attribute has a small set of unique values