Slide 12:
Here is a really short overview of the machine learning library. The MLlib library contains algorithms and utilities for classification, regression, clustering, collaborative filtering and dimensionality reduction. Essentially, you would use this for specific machine learning use cases that requires these algorithms. In the lab exercise, you will use the clustering K-Means algorithm on a set of taxi drop off points to figure out potentially where the best place to hail a cab would be.

Slide 13:
The GraphX is another library that sits on top of the Spark Core. It is basically a graph processing library which can used for social networks and language modeling. Graph data and the requirement for graph parallel systems is becoming more common, which is why the GraphX library was developed. Specific scenarios would not be efficient if it is processed using the data-parallel model. A need for the graph-parallel model is introduced with new graph-parallel systems like Giraph and GraphLab to efficiently execute graph algorithms much faster than general data-parallel systems.

There are new inherent challenges that comes with graph computations, such as constructing the graph, modifying its structure, or expressing computations that span several graphs. As such, it is often necessary to move between table and graph views depending on the objective of the application and the business requirements.

The goal of GraphX is to optimize the process by making it easier to view data both as a graph and as collections, such as RDD, without data movement or duplication.

The lab exercise goes through an example of loading in a text file and creating a graph from it to find attributes of the top users.

Summary:
Having completed this lesson, you should be able to understand and use the various Spark libraries.

Next steps:
Complete lab exercise #4, creating applications using Spark SQL, MLib, Spark Streaming, and GraphX and then proceed to the next lesson in this course.