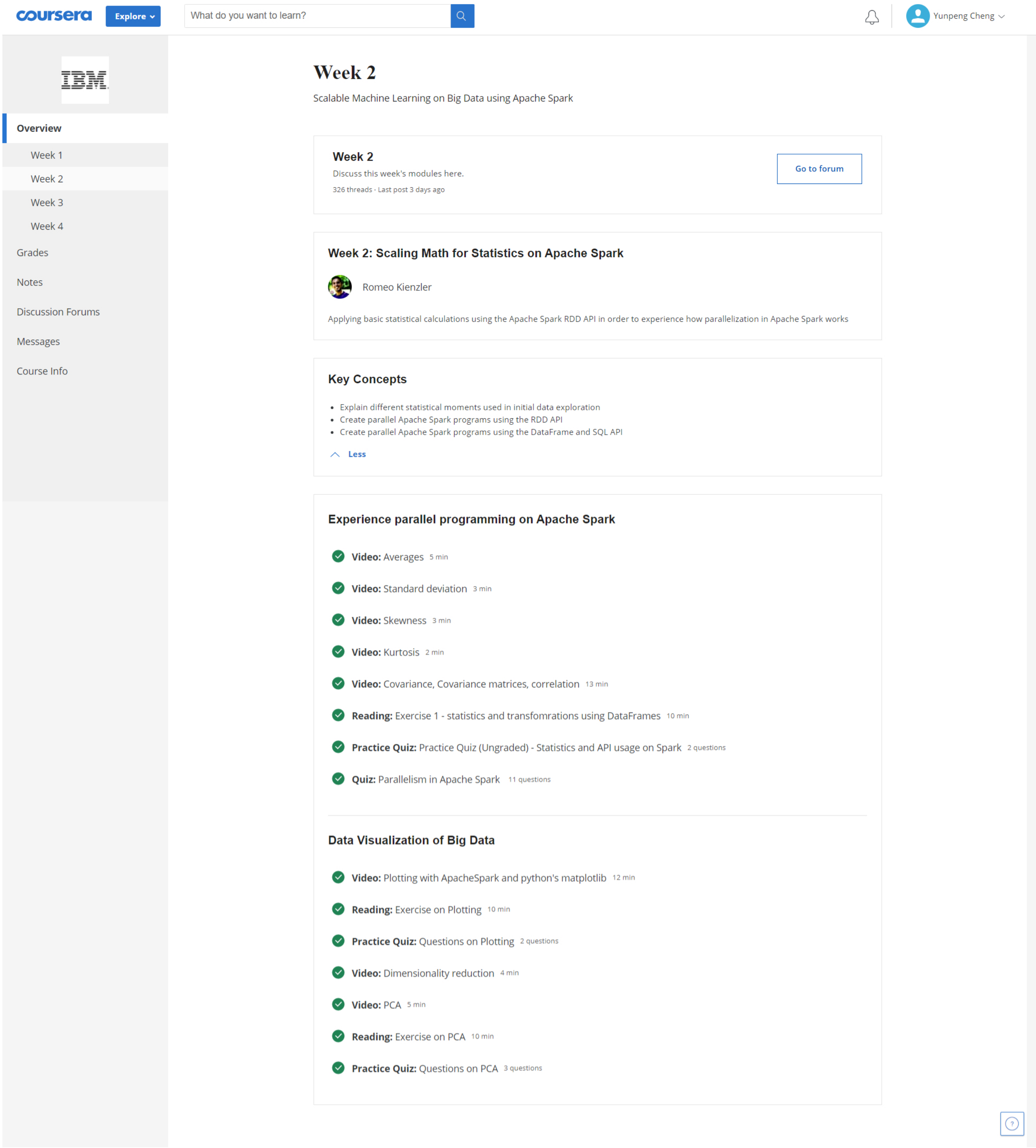

Week 2

Scalable Machine Learning on Big Data using Apache Spark

Week 2

Discuss this week's modules here.

326 threads · Last post 3 days ago

Week 2: Scaling Math for Statistics on Apache Spark

Romeo Kienzler

Applying basic statistical calculations using the Apache Spark RDD API in order to experience how parallelization in Apache Spark works

Key Concepts

- Explain different statistical moments used in initial data exploration
- Create parallel Apache Spark programs using the RDD API
- Create parallel Apache Spark programs using the DataFrame and SQL API

Experience parallel programming on Apache Spark

- ✓ **Video:** Averages 5 min
- ✓ **Video:** Standard deviation 3 min
- ✓ **Video:** Skewness 3 min
- ✓ **Video:** Kurtosis 2 min
- ✓ **Video:** Covariance, Covariance matrices, correlation 13 min
- ✓ **Reading:** Exercise 1 - statistics and transfomrations using DataFrames 10 min
- ✓ **Practice Quiz:** Practice Quiz (Ungraded) - Statistics and API usage on Spark 2 questions
- ✓ **Quiz:** Parallelism in Apache Spark 11 questions

Data Visualization of Big Data

- ✓ **Video:** Plotting with ApacheSpark and python's matplotlib 12 min
- ✓ **Reading:** Exercise on Plotting 10 min
- ✓ **Practice Quiz:** Questions on Plotting 2 questions
- ✓ **Video:** Dimensionality reduction 4 min
- ✓ **Video:** PCA 5 min
- ✓ **Reading:** Exercise on PCA 10 min
- ✓ **Practice Quiz:** Questions on PCA 3 questions