Основы Apache Spark

Код: EAS-017

Длительность: 26 ч.

Описание:

Apache Spark – один из самых распространенных инструментов, обеспечивающих гибкое масштабирование обработки данных в самых разных объемах. Если у вас есть кластер Spark, то достаточно один раз написать логику обработки данных на SQL с минимальным использованием кода на Python, Java или Scala и запустить приложение, независимо от того, хранится ли у вас 100 килобайт на одном узле или 100 терабайт на 100 узлах. Неизбежные сбои на узлах и сбои сетевой инфраструктуры в таких распределенных системах можно устранять с помощью того же Spark, перезапуская при необходимости неработающие процессы. Такие широкие возможности управления выполнением распределенных запросов в реляционных СУБД доступны либо при наличии большого бюджета (тогда как Apache Spark доступен бесплатно), либо при условии существенных трудозатрат на разработку.

Для эффективного использования всех преимуществ Spark недостаточно просто развернуть кластер и написать SQL запросы. Разработчики должны понимать, что происходит во внутренней структуре, в противном случае неизбежны неприятные сюрпризы с производительностью системы.

Этот тренинг ориентирован прежде всего на разработчиков и аналитиков данных, которые только начинают знакомство с фреймворком Spark, но не ограничивается базовыми понятиями. Будут рассмотрены различные способы оптимизации Spark не только в случае SQL-подобных запросов для табличных данных, но и для других типов данных, например, текстов, а также в случае взаимодействия с внешними системами, например, Cassandra и Greenplum.

В данной версии курса для практических упражнений используется язык Python. Это удобно для аналитиков данных, поскольку существенная часть упражнений выполняется в тетрадях Jupyter.

Цели:

Слушатели получат представление об основных концепциях и архитектуре Spark, научатся создавать табличные запросы, используя Spark SQL и DataFrame Python API, разрабатывать программы обработки данных как последовательности преобразований RDD, загружать данные для обработки Spark из систем JDBC, Kafka и Cassandra, а также сохранять полученные результаты во внешних хранилищах данных.

Целевая аудитория:

Разработчики, архитекторы, аналитики данных.

Предварительная подготовка – общее:

Базовые навыки программирования на языке Python. Базовые знания SQL..

Рекомендуемые дополнительные материалы, источники:

- Frampton M. Mastering Apache Spark. – Packt, 2015. – 476 p.
- Ryza S. et al. Advanced Analytics with Spark. – O'Reilly, 2015. – 261 p.
- Gupta S. Learning Real-time Processing with Spark Streaming. – Packt, 2015. – 271 p.
- Karau X. и др. Изучаем Spark: молниеносный анализ данных. – М.: ДМК Пресс, 2015. – 304 с.

Примечание:

Материалы курса представлены на английском языке.