Практический проект с использованием Hadoop

Код: EAS-023

Длительность: 8 ч.

Описание:

Рассматриваются основы Apache Hadoop и методы разработки приложений, обрабатывающих данные на его основе.

Участники познакомятся с проектами, составляющими экосистему Hadoop: HDFS, Hive, Spark, Sqoop, Flume, Druid, Kafka. Основное соедержание курса – разработка проекта, включающего загрузку, подготовку и извлечение данных.

Цели:

- понимать ключевые концепции и архитектуру Hadoop;
- получить представление об экосистеме, сложившейся вокруг Hadoop, и ее ключевых компонентах;
- уметь записывать и читать данные в/из HDFS, готовить файлы данных в HDFS для использования в SQL-запросах;
- уметь использовать Hive и Spark SQL для SQL-запросов
- уметь использовать Sqoop и Flume для загрузки данных.

Разбираемые темы:

1. Data storage and processing provisioning:

- HDFS Cluster;
- YARN Cluster;
- YARN-based Spark Cluster;
- Druid Cluster;
- Hive, Metastore, HCatalog;
- Sqoop;
- Flume.

2. Data ingestion:

- Model and create Hive data warehouse;
- Acquire user accounts data (source: Oracle or other RDBMS): Sqoop;
- Continuously acquire user activity streams (sources: log files in CSV, Kafka topics): Flume.

3. Data cleaning and transformation:

- Develop ETL in Hive;
- Develop ETL in Spark SQL.

4. Alerting: Develop near-real time outlier detection in Spark Streaming.
5. Analytics: Discover user segmentation model using Spark ML.
6. Real-time analytics: Design Druid-based OLAP cube for pre-defined reports.
7. Data storage and processing provisioning:

- HDFS Cluster;
- YARN Cluster;
- YARN-based Spark Cluster;
- Druid Cluster;
- Hive, Metastore, HCatalog;
- Sqoop;
- Flume.

8. Data ingestion:

- Model and create Hive data warehouse;
- Acquire user accounts data (source: Oracle or other RDBMS): Sqoop;
- Continuously acquire user activity streams (sources: log files in CSV, Kafka topics): Flume.

9. Data cleaning and transformation:

- Develop ETL in Hive;
- Develop ETL in Spark SQL.

10. Alerting: Develop near-real time outlier detection in Spark Streaming.
11. Analytics: Discover user segmentation model using Spark ML.
12. Real-time analytics: Design Druid-based OLAP cube for pre-defined reports.

Целевая аудитория:

Разработчики, архитекторы, разработчики баз данных.

Предварительная подготовка – общее:

- Базовые навыки программирования на Java.
- Умение работать в командной оболочке Unix/Linux (bash).
- Опыт работы с базами данных желателен, но не обязателен.

Примечание:

Материалы курса представлены на английском языке.