



BIG DATA 10.1 FOR DEVELOPERS

Слушатели данного тренинга смогут лучше понять работу Informatica Big Data Management и узнать, как правильно оптимизировать решения, созданные на его базе, для организации эффективных процессов загрузки хранилищ данных. Они смогут использовать механизмы переноса логики обработки на платформу Hadoop, а также смогут изучить технику работы с NoSQL базами данных и процессы сложных обработок файлов.

В рамках тренинга будут рассмотрены необходимые теоретические знания по продукту, а также разработан набор лабораторных работ, помогающих слушателю лучше понять принципы работы продукта.

После успешного завершения тренинга слушатели смогут:

- Работать с “Big Data”;
- Переносить логику вычислений на платформу Hadoop;
- Мигрировать маппинги PowerCenter в Big Data Management и исполнять их на платформе Hadoop;
- Мигрировать и обрабатывать данные на платформе Hadoop, используя SQOOP и SQL to Mapping;
- Понимать место Informatica в архитектуре Hadoop;
- Обрабатывать данные на платформе Hadoop, используя Informatica polyglot computing;
- Применять Informatica Smart Executor;
- Использовать средства мониторинга для поиска ошибок;
- Работать со сложными данными, такими как JSON, AVRO и Parquet;
- Понимать, как работает Informatica с NoSQL данными.

Тренинг, в первую очередь, будет интересен следующим категориям специалистов:

- разработчикам, непосредственно работающим в проектах с Big Data;
- аналитикам, архитекторам и руководителям проектов, участвующим в проектах с Big Data;
- кроме того, к дополнительной целевой аудитории также относятся квалифицированные специалисты, желающие понять суть работы с Big Data и повысить свой профессиональный уровень

Продолжительность тренинга составляет 3 дня (24 академических часа)

Каждый посетитель семинара должен иметь при себе ноутбук следующей конфигурации:

Компьютер PC или Mac: <ul style="list-style-type: none">• не ниже CPU: 1.3 GHz• не менее 2 GB RAM• Размер экрана не менее 1024 x 768 или выше	Операционная система: <ul style="list-style-type: none">• Windows XP и выше• Mac OS X• Linux
Поддерживаемые браузеры (не ниже): Internet Explorer 10.0, Firefox 15, Safari 6.0, Opera 12.0, Chrome 16	

Для наиболее полного усвоения программы участникам тренинга предварительно рекомендуется посетить тренинг: Informatica Developer Tool 10.1 Big Data Management. Также



предполагается, что участники тренинга имеют представление об алгоритмическом программировании и владеют английским языком на уровне чтения технической документации.

Программа тренинга по курсу BIG DATA 10.1 FOR DEVELOPERS

Модуль 1: Обзор курса Big Data Integration

- *Вступление*
- *Программа тренинга*
- *Доступ к учебным средам*
- *Дополнительные тренинги*

Модуль 2: Основы Big Data

- *Что такое Big Data?*
- *Концепция Hadoop*
- *Основные компоненты архитектуры Hadoop*
- *The Hadoop Distributed File System (HDFS)*
- *Назначение Name Node & Secondary Name Node*
- *MapReduce*
- *“Yet Another Resource Manager” (YARN) (MapReduce Version 2)*

Модуль 3: Загрузка данных в хранилище

- *Основные проблемы в традиционных хранилищах данных*
- *Требования к оптимальному хранилищу данных*
- *Процессы загрузки данных*

Модуль 4: Загрузка и выгрузка данных

- *PowerCenter Reuse Reports*
- *Импорт маппингов PowerCenter в Developer*
- *SQOOP*
- *SQL to Mapping capability*
- *Партиции и параллельная загрузка*

Модуль 5: Big Data Management

- *Архитектура*
- *Мир Big Data*
- *Создай один раз запускай везде*
- *The Informatica abstraction layer*
- *Polyglot computing*
- *The Smart Executor*
- *Open source и инновации*
- *Архитектура подключений*
- *Использование сторонних приложений*

Модуль 6: Informatica Polyglot

- *Управление в Hadoop*
- *Hive MR/Tez*



- *Blaze*
- *Spark*
- *Native*
- *The Smart Executor*

Модуль 7: Маппинги, мониторинг, поиск ошибок

- *Конфигурирование и запуск маппингов в Native и Hadoop средах*
- *Исполнение планов*
- *Мониторинг маппингов*
- *Поиск ошибок в маппингах*
- *Просмотр результатов*

Модуль 8: проблемы Hadoop Data Integration и оптимизация производительности

- *Описание проблем с запуском процессов в Hadoop*
- *Тюнинг процессов Big Data Management*
- *Оптимизация среды Hive*
 - *Оптимизация на уровне маппинга*
 - *Оптимизация на уровне сервиса DIS*
 - *Оптимизация на уровне кластера*
 - *Оптимизация на уровне кластера Hadoop*
- *Рекомендации*

Модуль 9: Качество данных в Hadoop

- *Процесс Data Quality*
- *Исследование потенциала данных*
- *Совместная работа разных ролей проекта*
- *Модификация, управление и мониторинг Data Quality*
- *Self Service Data Quality*
- *Запуск маппингов Data Quality на Hadoop*

Модуль 10: Обработка сложных файлов

- *Чтение сложных файлов*
- *Трансформация Data Processor*
- *Запись сложных файлов*
- *Производительность: Партиционирование*
- *Обработка и парсинг: Avro, Parquet, JSON, and XML file*
- *Договоренности при работе с Data Processor Transformation*

Модуль 11: Работа с NoSQL БД

- *CAP Theorem*
- *HBase*
- *MongoDB*
- *Cassandra*