Big Data & Analytics Fundamentals

Instructor: Mike Ferguson
Duration: 2 Days

OVERVIEW

This two-day workshop is aimed at getting Data Scientists, Data Warehousing and BI professionals up to scratch on Big Data technologies such as Hadoop, Storm, Spark, Flink, SQL on Hadoop, NoSQL DBMSs and Multi-Platform Analytics. What is Big Data? How can you make use of it? How does it fit within a traditional analytical environment? What skills do you need to develop for Big Data Analytics? All of these questions are addressed in this new knowledge packed workshop.

AUDIENCE

IT directors, CIO's, IT Managers, BI Managers, data warehousing professionals, data scientists, enterprise architects, data architects

LEARNING OBJECTIVES

Attendees to this seminar will learn:

- What Big Data is
- How to clearly understand business use cases for different Big Data technologies
- How to make use of Big Data to deliver business value
- How to set up and organise Big Data projects including skills
- How Big Data creates several new types of analytical workload
- Big Data technology platforms like Hadoop, Spark and NoSQL databases
- Big Data advanced analytical techniques and tools
- How to analyse un-modelled, multi-structured data using Hadoop, MapReduce & Spark
- How to integrate Big Data with traditional data warehouses and BI systems

MODULE 1: AN INTRODUCTION TO BIG DATA

This session defines big data and looks at business reasons for wanting to make use of this new area of technology. It looks at Big Data use cases and what the difference is between traditional BI and Data Warehousing versus Big Data

- What is Big Data?
- Types of Big Data
- Why analyse Big Data?
- Industry use cases - Popular big data analytic applications
- Data Warehousing and BI Versus Big Data
- The need to analyse new more complex data sources
- New skills - Data Science
- Types of Big Data analytical workloads
- Analysing high velocity streaming data at scale
- Exploratory analysis of multi-structured data
- Complex analysis of structured data
- Graph analytics
- Challenges when managing and analysing big data
- Architecture: The Extended Analytical Ecosystem
- Logical Data Warehouse
- Key components in a Big Data Analytics environment
- Preserving existing BI/DW investments

MODULE 2: BIG DATA PLATFORMS AND STORAGE OPTIONS

This session looks at platforms and data storage options for big data analytics

- The new multi-platform analytical ecosystem
- Beyond the data warehouse – Hadoop, Apache Spark, Apache Flink, NoSQL DBMSs, Analytical RDBMSs, NewSQL DBMSs
- NoSQL DBMSs
  - ○ Key Value stores, Document DBMSs, Column Family DBMSs and Graph databases
- An introduction to Hadoop and the Hadoop Stack
- HDFS, Kudu, MapReduce, Yarn, Pig, Hive
- Apache Spark Framework
- SQL on Hadoop options
  - ○ Impala, Hive, SparkSQL, HawQ, IBM Big SQL, MapR Apache Drill, Citus Data, Jethro, Splice Machine, Oracle Big Data SQL, Presto, HP Vertica SQL on Hadoop, Teradata QueryGrid
- The Big Data Marketplace
  - ○ Hadoop distributions – Cloudera, Hortonworks, MapR, IBM BigInsights and Open Platform with Apache Hadoop, Microsoft HD Insight
  - ○ Big Data Appliances – Oracle Big Data Appliance, IBM PureData System for Hadoop, HPE Big Data Platform, Teradata Aster Discovery Server
  - ○ NoSQL databases, e.g. Basho Riak, MongoDB, DataStax (Cassandra), Neo4j, Cray
- The Cloud deployment option – Amazon Elastic MapReduce, Cloudera Altus, Google DataProc and BigQuery, Microsoft Azure (HDInsight, Data Lake & Data Factory), IBM Bluemix, Qubole, Oracle Analytics Cloud, SAP Altiscale Data Cloud

MODULE 3: INTEGRATING BIG DATA ANALYTICS INTO THE ENTERPRISE

This session looks at how new Big Data platforms can be integrated with traditional Data Warehouses and Data Marts. It looks at stream processing, Hadoop, NoSQL databases, Data Warehouse appliances and shows how to put them together in an end-to-end architecture to maximise business value from Big Data

- Integrating Big Data platforms with traditional DW/BI environments – what's involved
- Integrating stream processing with Hadoop and Analytical DW Appliances
- Integrating Hadoop with DW Appliances and Enterprise Data Warehouses
- Tying together front-end tools

- Options for implementing multi-platform analytics
- Cross-platform analytical workflows
- The role of Data Virtualisation in a Big Data environment
- Creating a multi-platform analytical ecosystem and logical data warehouse architecture
- Multi-platform optimisation

MODULE 4: BIG DATA INTEGRATION AND GOVERNANCE IN A MULTI-PLATFORM ANALYTICAL ENVIRONMENT

This session will look at the challenge of integrating and governing Big Data and the unique issues it raises. How do you deal with very large data volumes and different varieties of data? How does loading data into Hadoop differ from loading data into analytical relational databases? What about NoSQL databases? How should low-latency data be handled? Topics that will be covered include:

- Types of Big Data
- Connecting to Big Data sources, e.g. web logs, clickstream, sensor data, and multi-structured content
- Data warehouse ETL offload
- Loading Big Data – what's different about loading HDFS, Hive & NoSQL Vs analytical relational databases
- Streaming data ingest – Kafka, Streamsets, Apache Apex, Flink,
- Change data capture – what's possible, e.g. Attunity Replicate & Kafka
- Parsing unstructured data
- Options - ETL tools Vs Pig Vs self-service DI/DQ
- ELT and Data Quality processing at scale on Hadoop and Spark
- Governing data in a Data Science environment
- Joined up analytical processing from ETL to analytical workflows
- The impact of data scientist and end user self-service DQ/DI – Alteryx, Paxata, Trifacta, Tamr, MS Excel Power Query, MicroStrategy, Tableau
- Data governance in a big data environment
    - The role of a Data Lake
    - Data lake use cases
    - Establishing a Data Refinery
    - Investigative analysis - from ETL to analytical workflows when refining data
    - The importance of an Information catalog
    - Organising data in a data lake
    - Big data audit, protection and security – Dataguise, IBM Guardium, Imperva, Privitar, Protegrity
    - Supplying consistent data to other analytical platforms
    - Creating a queryable archive

MODULE 5: TOOLS AND TECHNIQUES FOR ANALYSING BIG DATA

This session looks at tools and techniques available to data scientists, business analysts and traditional DW/BI professionals to analyse big data. It looks how different types of developers and users can exploit Big Data platforms such as Hadoop, Spark and NoSQL databases using search, machine learning, text analytics, graph analysis from analytical applications, interactive notebooks, data mining and self-service BI tools as well as how vendors are making it easier to gain access both the NoSQL/Hadoop world and the Analytical RDBMS world by using data virtualisation.

- Data Science projects
- Creating Sandboxes for Data Science projects
- Options for analysing unstructured content – Search, Text analytics, custom Spark or MapReduce code and interactive notebook tools, data mining tools & self-service BI
- Text analysis and visualisation, Sentiment analysis and visualisation
- Clickstream analysis and visualisation
- Analysing big data using Spark and MapReduce Tools and applications for Hadoop, e.g. ClearStory Data, Datameer, Arcadia Data
- Exploratory graph analysis and visualisations
- Using search to analyse multi-structured data
    - Creating search indexes on multi-structured data
    - Building dashboards and reports on top of search engine indexed content
    - The integration of search with traditional BI platforms
    - Guided analysis using multi-faceted search
    - The marketplace: Apache Solr, Attivio, Cloudera Search, Connexica, DataRPM, HPE IDOL, IBI WebFocus Magnify, IBM Watson Explorer, LucidWorks, Microsoft, ThoughtSpot, Oracle Big Data Discovery, Quid, Splunk
- Analysing Big Data using Data Mining Tools, e.g. Knime, IBM SPSS, RapidMiner, SAS, TIBCO Statistica
- Analysing Big Data using Self-Service BI Tools and SQL on Hadoop, e.g. Excel, IBM Watson Analytics, MicroStrategy, Microsoft PowerBI, Qlik, SAP BusinessObjects Lumira, SAS Visual Analytics, Tableau, TIBCO Spotfire, Zoomdata
- Big data analytics – query performance enablers
- Managing stream computing in a Big Data environment
- Tools and techniques for streaming analytics

PRESENTER

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent analyst and consultant, he specialises in data management and analytics. With over 38 years of IT experience, Mike has consulted for dozens of companies. He has spoken at events all over the world and written numerous articles. Mike is Chairman of Big Data LDN – the fastest growing Big Data conference in Europe, and chairman of the CDO Exchange. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates. He teaches popular master classes in Analytics, Big Data, Data Governance & MDM, Data Warehouse Modernisation and Data Lake operations.

ONSITE TRAINING

All training can be given as onsite education, tailored to your company's requirements. For further details please contact us at training@intelligentbusiness.biz

INTELLIGENT
BUSINESS
STRATEGIES

Tel/Fax: (+44) 1625 520700
info@intelligentbusiness.biz
https://www.intelligentbusiness.biz