UTN.BA
UNIVERSIDAD TECNOLOGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

TEMARIOS

Temario: Curso Big Data – Base de Datos NoSQL MongoDB.

Unidad 1 - Surgimiento y Conceptualización de Bases de Datos

1. Valor de las Bases de Datos NoSQL
2. Cambios en la evolución tecnológica de las BD 3. Surgimiento de NoSQL. Necesidades que cubren.
4. Definicion de BD NoSQL
5. Tipos de bases de datos NoSQL: Key-value, documents, column-family, graph
6. Persistencia Políglota: definición y necesidad de soluciones
7. Cuando usar y cuándo no MongoDB

Unidad 2: creando, actualizando, borrando y consultando documentos MongoDB

1. Inserción
2. Removing
3. Updating
4. Querying

Unidad 3: Indices

1. Introducción al indexado
2. Tipos de índices
3. Administración de índices
4. Índices geoespeciales / full text search / Time to live
5. Explain() y hint()

Unidad 4: Agregación

1. Framework de Agregación
2. Operaciones de Pipeline
3. Comandos de Agregación
4. Map Reduce Operation

Unidad 5: Modelado y Consistencia de Datos

1. Agregado
2. Normalización versus Denormalización
3. Optimizaciones para Manipulación de Datos
4. Consistencia

Unidad 6: Replicación

1. Replica Set
2. Componentes de un Replica Set
3. Conectando el Replica Set a una aplicación
4. Configurando un Replica Set

Unidad 7: Particionamiento

1. Introducción
2. Configuración
3. Elección de una clave de Partición
4. Administración de Particiones
5. Balancer y Splitting

Unidad 8: Administración

1. Start y Stop
2. Monitoreo de MongoDB
3. Backups
4. Import / Export de archivos
5. mongostat / mongotop / estadísticas
6. Balancer Start/Stop/Cron por horarios y colecciones

Unidad 9: Seguridad en MongoDB

1. Creación de usuarios
2. Roles Built In
3. Roles User-Defined 4. Seguridad en Clusters.

Unidad 10: Profiling/Auditoria en MongoDB / gridFS

1. Utilización de Profiling para detección de queries lento
2. Auditoria (Teórica - Enterprise Edition)
3. Creación de una GridFS / carga, consulta y recuperación de archivos

Unidad 11: Programación de Funciones en Java Scripts

1. Introducción a Java Scripts
2. Ejecución de bloques anónimos en JS
3. Creación y ejecución de funciones "user defined"
4. Creación de secuencias autonuméricas (utilizando funciones, colecciones y comando findAndModify)

Temario: Curso Big Data – Apache Hadoop.

Unidad 1: Surgimiento y Conceptualización de la Plataforma

1. Introducción a Big Data
2. Introducción a Apache Hadoop y sus componentes

Unidad 2: HDFS – Hadoop Distributed File System

1. Introducción a HDFS
2. Comandos de HDFS
3. Arquitectura de HDFS – Namenodes / Datanodes
4. Replicación y Alta Disponibilidad
5. Fail Tolerance

Unidad 3: HIVE - HQL

1. Introducción a HIVE
2. Arquitectura de Hive – MetaStore / Hive Server / Datawarehouse
3. Comandos de DDL – Data Definition Language 4. Comandos de DML – Data Manipulation Language 5. Particionamiento y Clustering de Datos.
6. Performance

Unidad 4: SQOOP

1. Introducción a SQOOP
2. Comandos para exportar datos desde Hadoop a un Motor de BD Relacional.
3. Comandos para importar datos desde un Motor de BD Relacional hacia Hadoop.
4. Interacción con Hadoop desde Herramienta ETL Pentaho Data Integrator.

Unidad 5: YARN

1. Introducción a YARN
2. Arquitectura de YARN – ResourceManager / NodeManagers / Containers
3. Colas de procesos.

UNIDAD 6: Motores de Procesamiento Distribuido

1. Introducción a Map Reduce como Modelo Conceptual.
2. Motor de Procesamiento Map / Reduce
3. Motor de procesamiento Tez 4. Motor de Procesamiento Spark

5.

UNIDAD 7: Lenguajes para procesamiento distribuido

1. PIG LATIN.
2. SCALA

UNIDAD 8: Procesamiento de Streaming

1. Concepto de Arquitectura Lambda
2. Introducción a KAFKA
3. Arquitectura de KAFKA – Brokers / Consumers / Producers
4. KAFKA Manager
5. Introducción a FLUME
6. Arquitectura de FLUME – Sources / Channels / Sinks
7. Introducción a Nifi
8. Configuración de Proceso de Ingesta de Datos de Streaming desde Twitter hacia Hadoop y MongoDB.

UNIDAD 9: Seguridad / Auditoria / Data Governance

1. Seguridad Básica
2. Introducción a KNOX
3. Introducción a RANGER
4. Arquitectura de RANGER
5. Introducción a ATLAS

Temario: Curso Big Data – Bases de Datos NoSQL Cassandra y Neo4J.

Unidad 1: Surgimiento y Conceptualización de Bases de Datos

1. Valor de las Bases de Datos NoSQL
2. Cambios en la evolución tecnológica de las BD
3. Surgimiento de NoSQL. Necesidades que cubren.
4. Definicion de BD NoSQL
5. Tipos de bases de datos NoSQL: Key-value, documents, column-family, graph
6. Persistencia Políglota: definición y necesidad de soluciones
7. Cuándo usar y cuándo no bases de datos basadas en Grafos y basadas en Familia de Columnas.

Unidad 2 : Conceptos básicos, Instalación y Configuración de Cassandra

1. Instalación de Cassandra

2. Identificación de key files y carpetas
3. Configuración de un Nodo de Cassadra
4. Iniciación y Parada de una Instancia Cassandra
5. Definición e identificación de conceptos principales (Cluster / KeySpace / ColumnFamily / RowKey).

Unidad 3 : insertando, borrando y consultando datos en Cassandra

1. Introducción a CQL
2. Creación de Estructuras
3. Introducción a Claves Primarias y claves compuestas.
4. Inserción de Datos
5. Borrado de Datos
6. Ejecución de consultas

Unidad 4: Arquitectura de Cassandra

1. Introducción a la Arquitectura basada en nodos de Cassandra
2. Introducción al proceso de particionamiento
3. Introducción al manejo de nodos virtuales
4. Entendiendo la replicación
5. Understanding hinted handoff
6. Introducción a Niveles de Consistencia
7. Entendiendo la Consistencia configurable

Unidad 5: Configuración de Replicación y Particionamiento

1. Estrategias de Replicación en Cassandra
2. Factor de Replicación
3. Creación de un Cluster
4. Introducción al manejo de nodos virtuales
5. Tipos de Particionamiento

Unidad 5: Indices y Seguridad en Casssandra

1. Creación de índices para consultas sobre columnas no rowkey
2. Utilización de índices
3. Estrategias de Seguridad a implementar

Unidad 6: Monitoreo y Mantenimiento de Casssandra

1. Monitoreo de Cassandra
2. Herramientas de Análisis

Unidad 7: Introducción a grafos y bases de datos basadas en grafos

1. Grafos y relaciones
2. Generalidades de las bases de datos basadas en grafos
3. Categorías de las bases de grafos
4. Casos de uso
5. Estructura interna de almacenamiento de Neo4j

Unidad 8: Operaciones sobre grafos en Neo4j

6. Nodos y relaciones en Cypher
7. Consultas generales
8. Creación de nodos y relaciones
9. Modificación de nodos y relaciones
10. Borrado de nodos y relaciones
11. Funciones de caminos
12. Índices

Unidad 9: Diseño de aplicaciones con Neo4j

1. Normalización y denormalización
2. Consistencia

3. Transacciones
4. Escalamiento

Unidad 10: Alta disponibilidad en Neo4j

1. Mecanismos para asegurar disponibilidad
2. Backup
3. HA Cluster en Neo4j

Unidad 11: Algoritmos de recorridos de grafos

1. Algoritmos clásicos de recorridos de grafos
2. Aplicación de algoritmos de grafos a bases de datos con Neo4j

Temario: Curso Big Data – Base de Datos NoSQL Redis.

Unidad 1: Surgimiento y Conceptualización de Bases de Datos

8. Valor de las Bases de Datos NoSQL
9. Cambios en la evolución tecnológica de las BD
10. Surgimiento de NoSQL. Necesidades que cubren.
11. Definición de BD NoSQL
12. Tipos de bases de datos NoSQL: Key-value, documents, column-family, graph
13. Persistencia Políglota: definición y necesidad de soluciones
14. Cuándo usar y cuándo no bases de datos Key-values.

Unidad 2: Estructuras de datos Básicas en Redis

7. Strings
8. Counters
9. HyperLogLog
10. Hashes
11. Lists
12. Sets

13. Sorted Sets

14. Keys & TTLs

15. Performance y notación Big-O

16. Operaciones con claves

Unidad 3: Estructuras avanzadas & Arquitectura

1. Geospatial
2. Publish / Subscribe
3. Usando Redis desde un lenguaje (nodeJS/Java)
4. Arquitectura: Persistencia / Arquitectura: Replicación & Particionamiento. (45 min basado en lo que tardamos en NOSQL)

Unidad 4: Streams

1. Introduction to Messaging, Streams, and Distributed Systems
2. Redis Streams
3. The Producer
4. The Consumer

Unidad 5: Extendiendo Redis

1. Transacciones y manejo de / Scripts LUA / Redis Modules (30 min basado en lo tenemos de NOSQL)
2. A module: Redis Search (Lo pongo por si nos quedamos corto de tiempo)

Temario: Curso Big Data – Elastic Stack

Unidad 1 - Introducción a Elastic Stack

1. Casos de aplicación de negocio
2. Funcionalidades de las herramientas del stack
3. Instalación del Stack Completo
4. Configuración recomendada
5. Cuando usar y cuándo no Elastic Stack

Unidad 2: Recolectando Datos por medio de beats

1. Filebeat
2. Packetbeat
3. Metricbeat

4. Heartbeat
5. Auditbeat
6. Winlogbeat
7. Configuración recomendada para Filebeat

Unidad 3: Logstash ingesta y transformación de datos

1. Introducción a la ingesta de datos
2. Arquitectura de Logstash
2. Inputs y Outputs
3. Utilización de Codecs
4. Filtros: Transformación a Datos Explotables
5. Monitorización

Unidad 4: Introducción a ElasticSearch

1. Arquitectura de un Cluster y Tipos de Nodos
2. Proceso de Indexación e Index Templates
3. Configuración del Motor de Búsqueda
4. Index Mapping

Unidad 5: Queries en ElasticSearch

1. Tipos de Queries
2. Match Queries y Term-Level Queries
3. Specialized Queries
4. Aggregation
5. Delete API
6. Update API
7. Bulk API

Unidad 6: Introducción a Kibana

1. Métodos de explotación de datos
2. Administración de Kibana
3. Administración de Índices
4. Discover

Unidad 7: Visualización de Datos en Kibana

1. Visualize
2. Tipos y Creación de visualizaciones
3. Dashboards
4. Canvas
5. Maps
6. Dev Tools
7. Monitoring

Temario: Curso Big Data – Pentaho Data Integration- Apache NiFi

Unidad 1: Business Intelligence

1. Definición de Sistemas de Información

2. Características, diferencias y similitudes de Sistemas OLTP y OLAP

3. Relación de sistemas de la información con la Inteligencia de Negocio (BI)
4. Definición, conceptos e historia de Business Intelligence
5. Diferencias entre conceptos DATOS – INFORMACIÓN - CONOCIMIENTO
6. Análisis OLAP - Multidimensional
7. Herramientas de visualización de reportes y tableros
8. ¿Por qué utilizar Business Intelligence?

Unidad 2 : BI Analítica, Funciones Analíticas. Data Warehouse y modelado

1. Concepto de BI Analítica, evolución y comparación con Inteligencia de Negocios tradicional
2. Presentación de Funciones Analíticas, Aggregate y Rank
3. Definición, características y objetivos de DataWareHouse
4. Alimentación DataWareHouse mediante procesos Ingesta (ETL)
5. Definición e identificación de conceptos principales (Desnormalización / Staging Area / Datamarts / Tecnología OLAP)
6. Enfoques de construcción de Datawarehouse
7. Objetivos y conceptos de modelado dimensional
8. Concepto de tabla de hechos y dimensiones asociadas
9. Definición de métricas e indicadores de progreso (KPI's)

Unidad 3 : Pentaho Data Integrator

1. Pentaho - Instalación y utilización
2. Definición y objetivos de la extracción, transformación y carga (ETL)
3. Introducción a las herramientas del suite de Pentaho
4. Herramienta ETL Pentaho Data Integrator y sus componentes
5. Conectividad, usos comunes y composición
6. Valores, metadatos, tipos de datos y métricas

Unidad 4: Aplicaciones de transformaciones y jobs en PDI

1. Introducción a transformaciones básicas
2. Introducción a componentes básicos

3. Extracción de datos de diversas fuentes
4. Limpieza de Datos de origen
5. Vuelco de información procesada en diferentes salidas
6. Modos de Ejecución, variables y parámetros

Unidad 5: Introducción a Apache NiFi

1. ¿Que es Apache NiFi?
2. ¿Porque usar Apache NiFi?
3. Conceptos principales y funcionalidades
4. Arquitectura y cluster

Unidad 6: Conceptos Básicos

5. Apache NiFi - Instalación y utilización
6. Como analizar las colas de NiFi
7. Concepto de atributos en flujos de datos
8. Como consumir/ingestar archivos desde un directorio local o HDFS hacia otro directorio local o HDFS.
9. Como consumir datos desde un web server.

Unidad 7: Automatización y administración de flujos de datos

1. Como crear y reutilizar templates
2. Grupos de procesadores
3. Publicar en un tópico de Kafka y loguear en caso de error
4. Consumir de un tópico de Kafka y almacenar en MongoDB
5. Compresion y descompresion de archivos

a. Reconversión de gzip a bzip2 para grabar en HDFS

6. Repositorio de flujos y archive

Unidad 8: Procesadores de NiFi 1. NiFi

expression language

2. Procesadores:

    a. Jolt Transform
    b. Merge Record
    c. Evaluate JsonPath
    d. InferAvroSchema
3. Apache NiFi Registry

Temario: Curso Big Data – Programación Distribuida, Text Mining y Data Science Aplicada.

Unidad 1: Conceptos generales de big data

1. Repaso de conceptos de big data.
2. Repaso de la plataforma Hadoop para soluciones distribuidas.
3. Repaso de conceptos de bases de datos NoSQL y su relación a las arquitecturas distribuidas.
4. Rol del profesional de big data en las organizaciones (tales como data scientist, data engineer, etc).
5. Historia de la programación distribuida.

Unidad 2: Arquitectura distribuida

1. Introducción al concepto de cluster para aplicaciones relacionadas con problemáticas de big data.
2. Implementación de distribución y replicación de datos. Ventajas y desventajas.
3. Problemáticas y técnicas de escalabilidad en big data. Planificación a futuro.
4. Introducción a arquitectura lambda

Unidad 3: Machine learning y algoritmos de data mining

1. Problemáticas de clasificación y algoritmos tales como Naive Bayes, Decision Tree, etc.
2. Problemáticas de regresión y su diferencia con clasificación.
3. Problemáticas de clustering y algoritmos tales como k-means y variantes.
4. Reducción dimensional, usos y aplicaciones.

Unidad 4: Lenguajes y Motores de procesamiento

1. Nociones generales de Java y Python para ambientes distribuidos.
2. Primitivas fundamentales de Spark y Map reduce.
3. Procesamiento real time vs procesamiento batch en herramientas distribuidas.

Unidad 5: Validación de resultados y testing

1. Técnicas de validación como (por ejemplo, cross validation, split validation).
2. Interpretación y visualización de resultados. 3. Técnicas de fabricación de variables artificiales.

TRABAJOS PRÁCTICOS

Listado de Trabajos Prácticos Curso Big Data - Apache Hadoop.

TP 1: Ejecución de HDFS Comandos desde el SO
TP 2: HDFS Comandos desde Hive Views
TP 3: HIVE HQL – DDL
TP 4: HIVE HQL – DML
TP 5: Ingesta de Datos con SQOOP – ETL con Pentaho DI
TP 6: Programación en PIG
TP 7: Utilización de Kafka y Flume para Publisher/Subscribe. Consultas de Zookeeper.
TP 8: Realizar CRUD en la BD Hbase
TP 9: Programación en Scala sobre Spark y Spark Streaming.
TP10: Utilización de NiFi (Hortonworks DataFlow), Kafka y Flume para procesar datos de Twitter.
TP11: Utilización de Zeppelin, armado de notes y párrafos.

Listado de Trabajos Prácticos Curso Big Data - NoSQL MongoDB.

TP 1: Consultas sobre MongoDB, creación de BD y Colecciones.
TP 2: Consultas Avanzadas, modificación, borrado e inserción de Documentos.
TP 3: Manejo de Arrays, comandos sabe y findAndModify
TP 4: Creación y manejo de índices.
TP 5: Desarrollo de consultas con el Aggregation Framework.
TP 6: Programación en JavaScript, desarrollo de funciones y secuencias.
TP 7: Configuración de Replicación de Datos y consistencia en lecturas y escrituras.
TP 8: Configuración de Distribución/Sharding
TP 9: Práctica integrada de replicación y sharding.
TP10: Utilización del Profiler y realización de configuraciones

Listado de Trabajos Prácticos Curso Big Data – NoSQL Cassandra/Neo4J.

TP 1: Modelado y Consultas Básicas sobre Cassandra.
TP 2: Consultas avanzadas, borrado, modificación y consultas.
TP 3: Configuración de Replicación y Distribución de Datos d
TP 4: Creación de Indices en Cassandra
TP 5: Práctica de Monitoreo
TP 6: Consultas sobre Neo4J en Cypher
TP 7: Consultas Avanzadas, modificación, borrado e inserción de Datos.
TP 8: Creación y manejo de índices.
TP 9: Replicación de Datos y Seguridad.
TP10: Backup, Monitoreo y Performance.

Listado de Trabajos Prácticos Curso Big Data – NoSQL Redis.

TP 1: Ejercicio de CRUD y Tipos de Datos.
TP 2: Ejercicios de CRUD Avanzados.
TP 3: Ejercicios de Monitoreo y Arquitectura.

Listado de Trabajos Prácticos Curso Big Data – Elastic Stack.

TP 1: Ejercicios utilizando Beats.
TP 2: Ejercicios utilizando Logstash.
TP 3: Ejercicios con Elastic Search.
TP 4: Ejercicios con Kibana y Elastic Search.

Listado de Trabajos Prácticos Curso Big Data – Pentaho Data Integrator y Apache Nifi.

TP 1: Ejercicios con PDI Básicos.
TP 2: Ejercicios con PDI con mayor dificultad.
TP 3: Ejercicios con Nifi Básicos.
TP 4: Ejercicios con Nifi con mayor dificultad.

Listado de Trabajos Prácticos Programación Distribuida.

TP 1: Ejercicio de Programación Distribuida I.
TP 2: Ejercicio de Programación Distribuida II.
TP 3: Desarrollo de Algoritmos de Clasificación.
TP 4: Desarrollo de Algoritmos de Regresión.
TP 5: Desarrollo de Algoritmos en Java con Map Reduce.
TP 6: Desarrollo de Algorimos en Scala sobre Spark.