

Programa de curso DATOS 2

Objetivos del curso

Brindar al estudiante los conceptos esenciales relacionados a las bases de datos no relacionales (NoSQL) y a las estrategias de acceso, obtención y distribución a la información como índices, búsquedas, llaves, grafos entre otros.

Se incluirá como materia del curso Patrones de Arquitectura para Datos con enfoque a Sistemas Distribuidos, Escalabilidad, Alta Disponibilidad e Integración de Sistemas como: Command and Query Responsibility Segregation (CQRS), conceptos básicos dentro de Domain Driven Design (DDD) para datos y Arquitectura Lambda para modos de procesamiento batch o stream ("realtime") para la interacción con sistemas de almacenamiento de información.

Oportunidades del curso

Al finalizar este curso el Estudiante será capaz de conocer el panorama general necesario para la aplicación de diferentes estrategias de almacenamiento de información según los requerimientos en cada etapa del ciclo de vida de los datos, en su creación, almacenamiento, utilización, así como en su propósito. Así mismo poder identificar qué arquitectura de almacenamiento es la mejor opción para un proyecto en específico.

Reglamento del curso

- Se considera importante el comportamiento en general: la disciplina, puntualidad en el curso, respeto al catedrático y al resto de sus compañeros.
- Los alimentos en clase, uso de aparatos eléctricos (salvo computadora cuando sea necesaria) y llamadas telefónicas no se permiten dentro del salón de clase. Evite salir y entrar constantemente.
- **No se aceptará tarde la entrega de tareas ni se repondrán trabajos con ponderación. El mecanismo de entrega se definirá en cada asignación.**

- Solo se aceptan ausencias justificadas con notificación por correo electrónico al catedrático y es indispensable el 90% de asistencia al curso para optar a nota del examen final.

Descripción de 18 semanas 2 veces por semana sesiones de 1.20 - horas cada una

Tema	Tareas asignadas este día para la siguiente sesión
Definition and Introduction <ul style="list-style-type: none"> • NoSQL means • Context and History • RDBMS Challenges • Caching when is good and when is not too good • Reading and Reporting Databases • Motivations and Drivers for other data alternatives 	Investigación: qué opciones de replicación tiene y cómo se configuran en un RDBMS a seleccionar por el alumno.
LAB/RESEARCH <ul style="list-style-type: none"> • Configuración de Redis y JMeter 	
Definition and Introduction <ul style="list-style-type: none"> • Focusing on Events <ul style="list-style-type: none"> ○ Representing an Event ○ Using events to collaborate ○ Event Sourcing ○ Handling Events ○ Events and Commands • Messaging / Queues / Service Bus • CQRS - Command and Query Responsibility Segregation 	Investigación: Prepare una presentación donde incluya los siguientes temas: <ul style="list-style-type: none"> - Seleccione un service bus software y defienda por escrito por qué debe ser utilizado. - Seleccione un Message Queuing software y defienda por escrito por qué debe ser utilizado.
LAB/RESEARCH <ul style="list-style-type: none"> • Using an ORM • Decoupling Data Access • .Net Core Data Access (https://devblogs.microsoft.com/dotnet/net-core-data-access/) 	
Domain Driven Design <ul style="list-style-type: none"> • Model Expressed in Software • Life Cycle of a Domain Object 	Reading: Caching Strategies and How to Choose the Right One”
Data <ul style="list-style-type: none"> • Datasets • Different Data Types <ul style="list-style-type: none"> ○ Structured Data ○ Unstructured Data ○ Semi-structured Data ○ MetaData • Data Analysis • Data Analytics • Business Intelligence • Key Performance Indicators 	
LAB/RESEARCH <ul style="list-style-type: none"> • Kafka • ELK 	Reading:”Why you should use a graph database”

Scalability and Availability <ul style="list-style-type: none"> • Clusters • Files Systems and Distributed File Systems • Sharding • Replication • Combination of Strategies 	
NoSQL Data Store Types <ul style="list-style-type: none"> • Sorted Ordered Column-Oriented Stores • Key/Value Stores • Document Databases • Indexers (Lucene/Solr, Sphinx) • Graph Databases 	Reading: “Performance at Scale with Amazon ElastiCache”
LAB/RESEARCH <ul style="list-style-type: none"> • Kafka y ELK: mejores prácticas de configuración 	
Understanding the Storage Architecture <ul style="list-style-type: none"> • Column-Oriented Databases • HBase Distributed Storage Architecture • Document Store Internals • Key/Value Stores <ul style="list-style-type: none"> ○ Under the Hood Memcached ○ Redis Internals 	
Entregas en semana de Examen Parcial 1	
LAB/RESEARCH <ul style="list-style-type: none"> • Sphinx Standalone • Sphinx+MySQL 	
Understanding the Storage Architecture <ul style="list-style-type: none"> • Eventually Consistent Non-Relational Databases. <ul style="list-style-type: none"> ○ Consistent Hashing ○ Object Versioning ○ Gossip-Based Membership and Hinted Handoff ○ Merkle Trees 	
Managing Transactions and Data Integrity <ul style="list-style-type: none"> • RDBMS and ACID • Distributed ACID Systems <ul style="list-style-type: none"> ○ Consistency ○ Availability • Partition Tolerance • Upholding CAP <ul style="list-style-type: none"> ○ Compromising on Availability ○ Compromising on Partition Tolerance ○ Compromising on Consistency • Monotonicity 	Reading: “There is No Now ”
Managing Transactions and Data Integrity <ul style="list-style-type: none"> • CALM • CRDT • BASE • Consistency implementations <ul style="list-style-type: none"> ○ Distributed Consistency 	

○ Eventual Consistency	
LAB/RESEARCH Interaction with NoSQL Stores <ul style="list-style-type: none"> Performing CRUD Operations Querying NoSQL Stores Modifying Data Stores and Managing Evolution 	
Data Processing <ul style="list-style-type: none"> Parallel Data processing Distributed Data Processing Processing Workloads Batch Mode Real-time Mode Stream 	
A “new” paradigm for Big Data <ul style="list-style-type: none"> “NoSQL is not Panacea” What does a data system do? Properties of a (Big?) Data System Problems with fully incremental architectures Lambda Architecture Kappa Architecture 	
Building Data Lakes <ul style="list-style-type: none"> What is a Data Lake? Data Ingestion Methods 	
Building Data Lakes <ul style="list-style-type: none"> Data Cataloging Securing, Protecting and Managing Data Data Lake Monitoring Data Lake Optimization Querying Analytics 	
LAB/RESEARCH <ul style="list-style-type: none"> Lambda Architecture for Batch and Stream Processing https://d1.awsstatic.com/whitepapers/lambda-architecture-on-for-batch-aws.pdf 	
Performance Tuning <ul style="list-style-type: none"> Goals of Parallel Algorithms Influencing Equations Partitioning 	Reading: “Firing on All Cylinders_ The 2017 Big Data Landscape – Matt Turck”
Performance Tuning <ul style="list-style-type: none"> Scheduling in Heterogeneous Environments Additional Mapreduce Tuning 	
Indexing and Ordering Data Sets <ul style="list-style-type: none"> Essential Concepts Indexing and Ordering External Indexing Strategies 	Reading: “The enterprise data lake: Better integration and deeper analytics”
LAB/RESEARCH Cloud Data Storages <ul style="list-style-type: none"> Google App Engine Amazon SimpleDB, DynamoDB, Redshift 	

How Big Data Power Business <ul style="list-style-type: none"> • (Big) Data Business Model Maturity Index • Data Analytics Lifecycle • (Big) Data Value Creation Drivers • Envisioning Process • Traditional BI/DWH vs Modern scale-out data architecture 	Reading: "It's Time To Upgrade Business Intelligence To Systems Of Insight" by Boris Evelson
Entregas en Semana de Examen Parcial 2	
What Do We Mean by Data-Driven <ul style="list-style-type: none"> • Data Collection • Data Access • Reporting • Alerting • Hallmarks of Data-Drivenness • Analytics Maturity 	
Data Driven Culture <ul style="list-style-type: none"> • Open, Trusting Culture • Broad Data Literacy • Goals-First Culture • Inquisitive, Questioning Culture • Anti-HIPPO Culture • Data Leadership 	
Big Data <ul style="list-style-type: none"> • Definition • Characteristics <ul style="list-style-type: none"> ○ Volume ○ Velocity ○ Variety ○ Veracity ○ Value Properties of a (Big) Data System <ul style="list-style-type: none"> • Robustness and fault tolerance • Low latency reads and updates • Scalability • Generalization • Extensibility • Ad Hoc Queries • Minimal maintenance • debuggability 	
Visualizations	
13/11 -	
Entrega a finalizar semana de Examen final	

Bibliografía

1. **Professional NoSQL** - Shashank Tiwari - Wrox
2. **Domain Driven Design** - Tackling Complexity in the Heart of Software. - Eric Evans - Addison Wesley

3. **Big Data - Principles and best practices of scalable real-time data systems** - Nathan Marz
Manning
4. **Big Data Fundamentals - Concepts, Drivers & Techniques.** - Thomas Erl - *Prentice Hall*
5. **Big Data - Understanding How Data Powers Big Business** - Bill Schmarzo - *Wiley*
6. **Creating a Data-Driven Organization** - Carl Anderson - *O'Reilly*

Evaluación

Research/Asistencia/Labs	30%
2 Evaluaciones Parciales (Proyectos)	30%
Total Zona	60%
Evaluación Final (Proyecto Completado)	40%
Puntuación Total	100%