

Temas presentación final

La idea de las presentaciones es que en grupos de 4 personas exploren en más profundidad tópicos relacionados con este curso. En esta ronda de presentaciones los temas serán propuestos en torno a los temas vistos durante todo el curso. La idea es que investiguen en torno a un tema de los que se detallan a continuación y realicen un video de **entre 10 a 20 minutos** que van a subir a YouTube donde expongan los resultados de la investigación. En esta ronda de presentaciones, no deben entregar *feedback* a las presentaciones de los otros grupos.

1. Temas

Los temas para esta ronda son los siguientes.

1.1. Neo4J vs MongoDB vs SQL

En este tema se busca entender los casos de uso donde conviene usar Neo4J vs MongoDB vs SQL. Para esto, se espera que muestren casos concretos y/o papers que se hayan hecho la respecto. Además, en esta ocasión, se espera que diseñen un conjunto de experimentos con datos ficticios y consultas sobre estos datos. Se espera que midan el desempeño de cada uno de los sistemas en estas consultas.

1.2. Apache Spark: RDDs

En este tema se espera que definan y exploren en detalle el concepto de RDD (Resilient Distributed Datasets), que es la forma nativa que tiene Apache Spark para manejar datos distribuidos. Además, espera que además muestren ejemplos de cómo trabajar con este tipo de datos (por ejemplo, cómo cargar archivos, cómo ejecutar funciones, etc.). El punto de partida es la documentación oficial (<https://spark.apache.org/docs/latest/rdd-programming-guide.html>)

1.3. Apache Spark: Dataframes y Spark SQL

En este tema se espera que profundicen en los módulos de Spark que permiten cargar Dataframes y datos tabulares. Se espera que muestren ejemplos de cómo cargar datos que puedan ser trabajados como Dataframes y con Spark SQL. Además se espera que muestren ejemplos de las funciones más importantes para los Dataframes de Spark, y ejemplos de Spark SQL. Finalmente, se espera que discutan la relación entre los Dataframes de Spark y los Dataframes de Pandas.

1.4. PySpark

En clases vimos solamente Spark funcionando sobre Scala. La idea de este tema es que preparen un tutorial completo de PySpark, que es el módulo de Python para acceder a las funciones de Spark. Se espera ejemplos de carga y procesamiento de datos como los vistos en clases y otras funciones que consideren interesantes.

1.5. GraphX

En clases estudiamos cómo GraphX nos sirve para para ejecutar procedimientos basados en Pregel, sin embargo, no estudiamos en detalle las funcionalidades ya implementadas. En este tema se espera que hagan un tutorial del módulo GraphX considerando las funciones y herramientas que ya tiene implementadas (como por ejemplo PageRank, conteo de triángulos, entre otros). Además, se espera que muestren cómo cargar grafos a partir de archivos.

1.6. Pregel en Neo4J

En clases mostramos cómo usar Pregel en GraphX. La idea de este tema es que muestren cómo se puede usar Pregel en Neo4J, mostrando ejemplos como los vistos en clase pero para este sistema de bases de datos de grafos.

1.7. Procesamiento de eventos y Apache Kafka

En el curso no tocamos el tema de sistemas para procesamiento de eventos. Estos sistemas permiten procesar *streams* de datos en tiempo real. Uno de los sistemas más conocidos en este ámbito es Apache Kafka, una plataforma *open-source* para procesar eventos en *streaming*. La idea de este tema es introducir el tópico y además hacer un tutorial de Apache Kafka.

1.8. Apache Cassandra

En el curso no discutimos sobre el sistema Apache Cassandra, una base de datos muy popular para manejar grandes cantidades de datos. La idea de este tema es hacer un tutorial completo sobre Apache Cassandra, incluyendo instalación, cargar datos y hacer consultas. Además, se espera que discutan algunas otras funcionalidades de este sistema que consideren interesantes.

1.9. Otro tema

¿Tienes otro tema para esta ronda de presentaciones? ¡Publicítelo! Pueden proponer su propio tema como grupo, con previa autorización del profesor. La idea es que vaya en la misma línea de los otros temas planteados.

2. Detalles Académicos

Cada grupo deberar enviar un correo al profesor con su preferencia para dos temas hasta el jueves 17 de junio. Idealmente, no puede haber dos grupos que hagan el mismo tema, por lo que el tema de cada grupo se asignará por orden de llegada. El link del video se debe entregar por correo hasta el 11 de julio a las 23:59 horas, en donde el asunto del correo debe ser **[TICS575 - Video Presentación Final]**.