Cadre méthodologique des présentations

Les objectifs

1. Construire une idée panoramique chez l'élève ingénieur de l'écosystème des outils BI.
2. Apprendre à présenter et argumenter à propos d'une solution BI.
3. Savoir synthétiser plusieurs outils dans une conception globale d'un système BI.

Introduction

Les sujets proposés sont axés sur 4 thématiques qui se complètent dans un système BI:

- Les processus ETL/ELT et intégration des données
- Visualisation des données
- Entreposage des données
- Génération des données théoriques

Les 3 premières thématiques constituent les composants principaux d'une solution BI (Voir la figure 1). La dernière thématique est importante pour des objectifs comme les tests des dataproducts, validation d'hypothèses présupposées sur les data analysées, gestions de confidentialité et/ou rareté des data sources. Dans cette activité pratique, nous allons aborder un échantillon représentatif des outils existant dans l'écosystème BI. Voir le tableau 1.

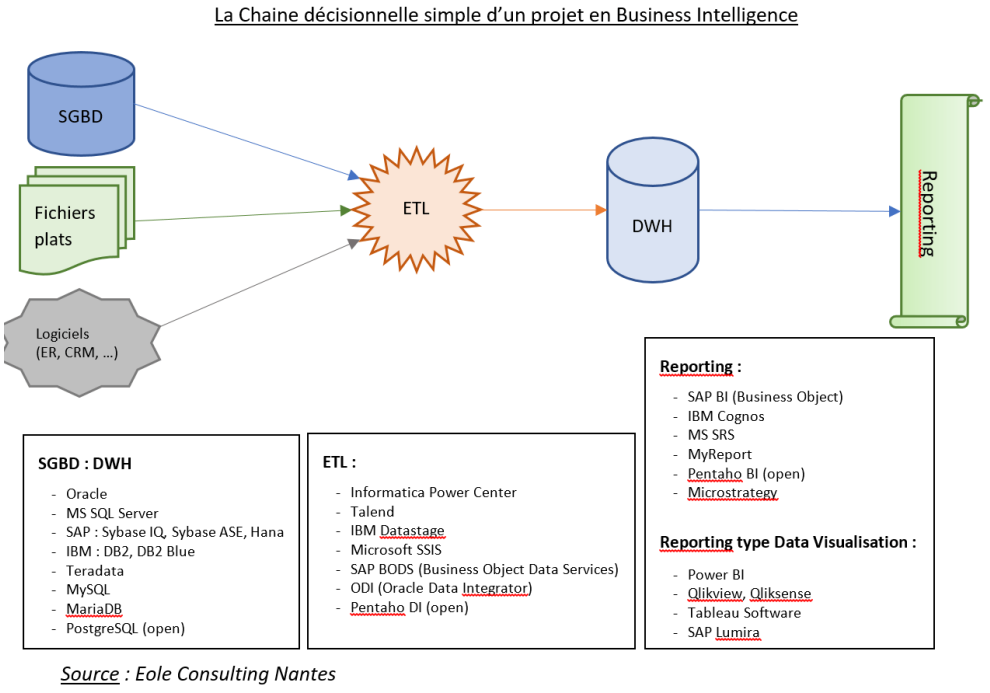

Figure 1. Un système BI simple.

Il faut noter que les outils mentionnés dans le tableau des sujets doivent être présentés dans le contexte d'une solution BI. Par exemple, l'outil Kafka peut être utilisé comme outils d'ingestion des données dans un contexte Big data (puisqu'il fournit les possibilités de traitement distribué), comme il peut être utilisé comme outil intégré dans un autre contexte d'une solution BI comme

outil de processus ETL. Une solution MySQL peut être implémentée dans un contexte de traitement de données OLTP pour représenter la circulation des informations de l'entreprise selon un système d'information préexistant, comme elle peut être implémenté dans un contexte de traitement de données OLAP pour implémenter un système de gestion de données conçu pour permettre et faciliter les activités de business intelligence (BI), en particulier l'activité de collecte, stockage et analyse.

Donc, les groupes des élèves ingénieurs sont amenés à présenter la thématique avec un esprit argumentatif en se référant aux questions directrices suivantes :

- Quelle est le contexte général d'apparition de l'outil présenté ?
- Comment cet outil est utilisé dans le contexte spécifique du BI ?
- Quel sont les composants généraux de l'outil, et surtout les composants en relation avec les solutions BI ?

Veuillez lire et respecter les consignes suivantes :

1. Durée de présentation : jusqu'à 45 min.

2. Présenter un outil dans un temps bref permet à l'élève ingénieur de simuler la pression à laquelle un data scientist peut faire face dans la vie professionnelle.

3. Adoptez un esprit argumentatif dans vos présentations. Imaginez que l'outils que vous allez présenter est votre solution et le professeur et les autres étudiants sont des stakeholders. Afin de normaliser le déroulement des exposés, les étudiants doivent suivre le plan suivant :
    - Une présentation générale du contexte d'apparition de l'outil.
    - Une présentation des fonctionnalités de l'outil. La concentration sera autours des fonctionnalités en relation avec les solutions BI.
    - Une présentation des composantes et l'architecture de l'outil. La concentration sera autours des fonctionnalités en relation avec les solutions BI. (Illustrer les concepts avec des figures ou des cartes conceptuelles)
    - Une présentation brève des informations commerciales de l'outil (open sources ? version actuelle ? la popularité parmi les concurrents? etc.)
    - Une démonstration sous forme de petit tutoriel à chaud ou avec des captures d'écrans expliquant des manipulations de l'outil.
    - Une présentation des comparaisons d'avantages et inconvénients avec d'autres outils concurrents sur le marché.

4. Gestion des rôles lors de la présentation :

    - Chargé de présentation technique : exécute les démonstrations sur PC, ou explique les captures d'écrans techniques etc.

- Chargé d'une section : présente le contenu d'une section.
- Modérateur de l'exposé : introduit la présentation, distribue les paroles aux différents intervenants, contrôle le temps etc

Tableau 1. Thématiques proposées.

Thématiques par axe	Axe des thématiques exposées
Généralités sur les outils pour ETL (usecases des : extracteurs de bases de données, webscrapers, API extractors, cloud extractors, document extractors ..)	Les processus ETL/ELT et intégration des données
Webscraping avec python / L'outil ScraperAPI	
Les processus ETL pour l'analyse des Documents (exemple : Optical Character Recognition (OCR))	
L'outil Talend Open Studio	
L'outil Pentaho Data Integration (PDI)	
Apache Kafka	
Apache Nifi	
Apache Aireflow	
L'outil Power bi / Tableau	Visualization des données
L'outil Looker Studio	
L'outil Microsoft SQL Server / SQL Server Integration Services (SSIS)	Entreposage des données
L'outil Snowflakes	
L'outil Teradata	
L'outil Mysql/oracle	
Génération des data synthétiques avec les GANs (Generative Adversarial Network) et data augmentation (CTGANs, DCGANs etc)	Génération des données théoriques
Génération des data synthétiques: Simulation des loi probabilistes avec R et/ou python	

L'ordre d'exécution des présentations seront comme suit :

(L'affectation des sujets aux groupes est aléatoire !)

1. Génération des data synthétiques: Simulation des loi probabilistes avec R et/ou ython(groupe 1)
2. Généralités sur les outils pour ETL (usecases des : extracteurs de bases de données, webscrapers, API extractors, cloud extractors, document extractors ..) (groupe 2)
3. L'outil Power bi / Tableau (groupe 3)
4. Apache Kafka (groupe 4)
5. Webscraping avec python / L'outil ScraperAPI (groupe 5)

6. L'outil Pentaho Data Integration (PDI) (groupe 6)
7. Apache Nifi (groupe 7)
8. Apache Aireflow (groupe 8)
9. Les processus ETL pour l'analyse des Documents (exemple : Optical Character Recognition (OCR)) (groupe 9)
10. L'outil Talend Open Studio (groupe 10)
11. L'outil Looker Studio (groupe 11)
12. Génération des data synthétiques avec les GANs (Generative Adversarial Network) et data augmentation (CTGANs, DCGANs etc) (groupe 12)
13. L'outil Microsoft SQL Server / SQL Server Integration Services (SSIS) (groupe 13)
14. L'outil MySql/Oracle (groupe 14)
15. L'outil Snowflakes (groupe 15)
16. L'outil Terdata (groupe 16)