

Prosit aller 1

L'équipe :

Animatrice : KADOUCHE Lydia

Secrétaire : HADDOUCHE Othmane

Scribe : OUCHAOUA Abdelmalek Ghiles

Gestionnaire : BESSAM Abdenour

1. Mots à définir :

1.1. Architecture BIGDATA :

Une architecture Big Data est conçue pour gérer l'ingestion, le traitement et l'analyse de données trop volumineuses ou complexes pour les systèmes de base de données traditionnels. Ces architectures sont utilisées pour :

- Stocker et traiter des données dans des volumes trop vastes pour une base de données traditionnelle.
- Transformer des données non structurées en vue d'une analyse et de la création de rapports.
- Capturer, traiter et analyser des flux de données indépendants en temps réel ou avec une faible latence.

1.2. BI (BDD) :

Power BI est un ensemble de services logiciels, d'applications et de connecteurs qui œuvrent ensemble pour transformer des sources de données disparates en informations visuelles immersives et interactives. Vos données peuvent être sous forme de feuille de calcul Excel ou de collection d'entrepôts de données hybrides locaux ou sur le cloud. Power BI vous permet de vous connecter facilement à vos sources de données, de visualiser et de découvrir ce qui est important, et de partager ces informations avec qui vous voulez.

2. Mots-clés :

- Modélisation des données
- Implémentation des données
- Exploitation des données
- Environnement de virtualisation

3. Contraintes :

- à 3 semaines

4. Contexte :

Archi a finalisé le projet CHU de son côté, de ce fait, nous avons pris le relais

5. Problématique :

Comment s'assurer que l'architecture BIGDATA proposée est correcte

6. Hypothèses :

6.1. L'architecture est correcte :

- Validation : Fausse

6.2. Il existe plusieurs architectures BIGDATA

- Validation : Vraie

7. Plan d'actions :

7.1. S'initier au BIGDATA et ses architectures :

7.1.1. Qu'est-ce que BIGDATA :

Ce terme signifie mégadonnées, grosses données ou aussi données massives, Ils désignent un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler.

7.1.2. D'où viennent ces données :

En effet nous sommes tous une source qui nourris ce domaine vaste qu'est la BIGDATA car nous procréons plus de 2 trillions d'octet tous les jours, en effet nous nourrissons ce domaine à chaque vidéo, message, enregistrement audio, utilisation de service en ligne

7.1.3. Les deux architectures utilisées en BIGDATA :

7.1.3.1. LAMBDA :

Ce modèle évolutif et tolérant aux incidents a été conçu pour gérer les mises à jour avec une faible latence.

L'autre avantage de cette architecture est le fait que les couches par lots et en temps réel sont totalement complémentaires et qu'une requête peut bénéficier des avantages des deux modes de traitements. La conception de l'architecture lambda est guidée par les contraintes suivantes :

- Passage à l'échelle : l'architecture proposée doit pouvoir passer à l'échelle de manière horizontale, c'est à dire en ajoutant des serveurs. Cette croissance doit se faire en garantissant la robustesse et la tolérance aux pannes des différents systèmes.
- Facilité de maintenance : les choix techniques réalisés ne doivent pas contraindre l'architecture à être figée dans le marbre. Il doit être aisé de déboguer et de modifier les applications exploitant cette architecture. Enfin, peu d'interventions manuelles doivent être nécessaires pour réaliser la maintenance des systèmes.

- Facilité d'exploitation des données : le but d'une architecture lambda n'est pas uniquement de stocker des données, mais également de les mettre à disposition d'autres applications pour les exploiter et en extraire de la valeur. Il doit être possible de réaliser des analyses personnalisées sur ces données de manière aisée.

7.1.3.2. KAPPA :

Partant du constat que la plupart des solutions effectuent un traitement en temps réel et un traitement par lots, Jay Kreps a créé une nouvelle architecture qui fusionne la couche batch et la couche real-time : l'architecture Kappa. Kappa n'est pas destinée à stocker les données Big Data mais uniquement à les traiter. Cette structure est de plus en plus utilisée dans les organisations car elle permet de mutualiser plusieurs couches et ainsi réduire la charge opérationnelle de l'architecte.

7.2. Etudier la modélisation des données (style BIGDATA) :

La modélisation des données (data modeling en anglais) est un processus de description de la structure, des associations, des relations et des contraintes relatives aux données disponibles. Elle sert à établir des normes et à coder des règles de gestion (modèles) des data dans l'organisation. Le data modeling fait partie intégrante de la phase de planification de tout déploiement analytique dans l'organisation ou de projet de Business Intelligence.

7.2.1. Les types de modélisation :

7.2.1.1. **La modélisation en étoile** : Les schémas en étoile sont très utilisés pour concevoir les Data Mart. La modélisation en étoile peut être directement associée à l'approche Top-Down que nous décrivons plus haut. La modélisation est obtenue par agrégation de tables de dimensions extraites du Data Warehouse.

7.2.1.2. **La modélisation en flocons** : Un modèle en flocon est un modèle pour lequel chaque dimension est représentée avec plusieurs tables. Il est donc plus normalisé (moins redondant) qu'un modèle en étoile.

7.2.1.3. **La modélisation en constellation** : C'est un schéma de représentation d'un modèle multidimensionnel. Il s'agit d'une collection de plusieurs tables de faits ayant des tables de dimensions communes. Il peut être considéré comme une collection de plusieurs schémas d'étoiles et, par conséquent, également connu sous le nom de schéma Galaxy. C'est l'un des schémas les plus utilisés pour la conception d'entrepôt de données et il est beaucoup plus complexe que le schéma en étoile et en flocon de neige. Pour les systèmes complexes, nous avons besoin de constellations de faits.

7.3. Analyser et étudier l'architecture ainsi que ses outils :

7.3.1. Quelques outils pour la Big Data :

- XPlenty
- Apache Hadoop
- Cloudera Distribution pour Hadoop
- Cassandra
- Knime
- Datawrapper
- MongoDB
- Lumify
- HPCC
- Tempête

7.4. Proposer l'architecture adéquat :

Les données Big Data seront transformées par Talend afin de les rendre exploitables et utilisables, par la suite ces données seront regroupées en plusieurs base de données grâce à HADOOP et enfin ces dernières seront traitées et utilisées grâce à Power BI.