Présentation SFE :

1. Big DATA :

Définition :

Ils désignent un ensemble très volumineux des données qu’aucun outil classique de gestion de base des données ou de gestion de l’information ne peut vraiment travailler.

Objectif :

Le Big Data se présente comme une solution dessinée pour permettre à tout le monde d’accéder en temps réel à des bases des données géantes. Il vise à proposer un choix aux solutions classiques de bases des données et d’analyse (plate-forme de Business Intelligence en serveur SQL…).

Marche du big data :

Dans son rapport, le monde du Big Data attend des investissements à hauteur de **130 milliards de dollars en 2016**, soit une hausse de 11,3% par rapport à 2015. Et en 2020, les analystes d’IDC prévoient des valeurs supérieures à 203 milliards de dollars.

Caractéristiques : Rapport page 10 -11

Domaine d’application : rapport figure 3 page 12

Comment gérer les big data :

Il existe plusieurs solutions qui peuvent entrer en jeu pour optimiser les temps de traitement sur des bases des données géantes à savoir les bases des données NoSQL (MongoDB, Cassandra ou Redis), les infrastructures du serveur pour la distribution des traitements sur les noeuds et le stockage des données en mémoire :

+La première solution permet d’implémenter les systèmes de stockage considérés comme plus performants que le traditionnel SQL pour l’analyse des données en masse (orienté clé/valeur, document, colonne ou graphe).

+La deuxième est aussi appelée le traitement massivement parallèle. Le Framework Hadoop en est un exemple. Celui-ci combine le système de fichiers distribué HDFS, la base NoSQL HBase et l’algorithme MapReduce.

1. Hadoop :

Introduction :

Hadoop est un projet open source d’Apache qui a commencé par Yahoo en 2006, il est un software librairie est un Framework qui permet de traiter les grands ensembles des données (big data).

Il est composé des modules qui travaillent ensemble pour créer le Framework Hadoop, les principaux modules sont :

+Hadoop Common

+Hadoop Distributed File System (HDFS)

+Hadoop YARN

+Hadoop MapReduce

FIGURE 4 page 15

Elle repose principalement sur HDFS (Hadoop File System), un système de fichier (file system), comme ceux qui gèrent les disques durs, capable de reconnaître et d’indexer les sources et formats des données SQL et NoSQL, donc d’offrir une vision unique d’un ensemble des données. En plus de HDFS pour le stockage des fichiers, Hadoop peut être aussi configurer pour utiliser S3 buckets ou Azure blobs comme des entrés.

Domaine d’application figure 5 pages 16

Avantage de Hadoop : titre page 17-18-19-20

Architecture Hadoop figure 6 pages 20 + un petit résumer des 4 composant

1. Spark :

Introduction :

Spark est un projet plus récent, initialement développé en 2012, à l'AMPLab à UC Berkeley. Il s'agit également d'un projet Apache de haut niveau axé sur le traitement des données en parallèle sur un cluster, mais la plus grande différence est qu'il fonctionne en mémoire.

Domaine d’application : page 31-32

Avantages : 32-33-34

Ecosysteme de spark : figure 18 page 35

Architecture de spark : partie f page 36 -37

1. Etude :

Introduction + resultat de l’etude

1. Différences :

Les pics li siftlk + chof yla kaynin chi w7din fdak chkel nit + dakchi li flpartie differences