Découvrir la plateforme Big Data

SOMMAIRE

La plateforme Big-Data de l'AP-HP constitue la brique technique principale de l'Entrepôt de Données de Santé.

OBJECTIFS PRINCIPAUX

Afin de répondre aux objectifs ambitieux que se donne l'AP-HP pour développer l'usage de ses données, et notamment permettre l'émergence et le développement de l'intelligence artificielle dans le domaine médical, la plateforme Big-Data a été mise en place et répond aux besoins techniques qui y sont liés, à savoir :

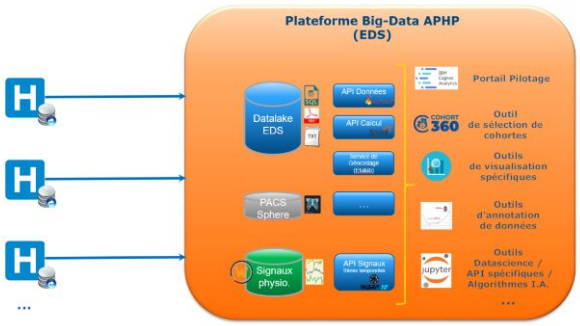
- Le stockage de données variées, complexes et volumineuses
- L'exploitation de ces données (ressources de calcul distribuées, GPUs ...)

Cette plateforme est hébergée par l'AP-HP dans un datacenter certifié pour l'hébergement de données de santé, les plus hauts niveaux de sécurité sont donc en place pour s'assurer à la fois de la sécurité des données et de la disponibilité de la plateforme.

ARCHITECTURE GLOBALE

La plateforme Big-Data supporte 4 briques principales :

- La récupération et la modélisation des données (ETLs)
- Le stockage des données (SQL, Hadoop, Solr)
- L'exposition des données et algorithmes (APIs)
- Les différents portails et outils web d'accès aux ressources et données



LA RÉCUPÉRATION ET LA MODÉLISATION DES DONNÉES (ETLS)

C'est la 1ère étape à la constitution d'un Entrepôt de Données de Santé. Il s'agit :

- De développer des flux (ETL) en se connectant aux +800 bases de données de l'AP-HP dédiées aux soins pour y récupérer les données
- De mettre ces données dans un même format, c'est la standardisation
- D'aligner les terminologies AP-HP vers des terminologies standard (LOINC, HL7-FHIR, NCBI...)

Un effort important est porté à l'intégration rapide des données cliniques (structurées et non structurées) produites dans les différents systèmes d'information hospitaliers (Dossier Patient Informatisé (*DPI*) ORBIS, logiciels historiques et de spécialité, données des moniteurs haute fréquence, données d'imagerie...) afin de pouvoir les mettre à disposition de tiers au travers d'outils spécifiques (logiciel I2B2, outil BI Cognos, outils propres de visualisation et de création de cohortes...) ou d'interfaces FHIR (API).

Un travail important de standardisation des données est réalisé afin de s'aligner sur les standards internationaux et assurer une interopérabilité maximale des données (OMOP, FHIR et les terminologies médicales de référence LOINC, CIM...).

LE STOCKAGE DES DONNÉES (SQL, HADOOP, SOLR)

Après la récupération des données et leur modélisation dans des standards internationaux, la seconde étape constitue à :

- Choisir le moyen de stockage de ces données pour leur future exploitation
- L'indexation des données dans des moteurs de recherche (SolR) pour permettre aux différents outils de chercher instantanément dans les données
- Executer des pipelines de post-traitement et d'enrichissement de la donnée, par exemple pour la pseudo-anonymisation des données

Un travail conséquent est réalisé afin de choisir la meilleure manière de stocker et d'exposer la donnée. Il s'agit dans un premier temps de réaliser une veille technologique, puis d'installer le système qui correspond le mieux à ces attentes, afin de stocker ces données dans des bases de données relationnelles classiques (PostgreSQL, MySQL), dans des systèmes de fichiers distribués (HDFS, GlusterFS...), dans des bases de données distribuées (Hive, HBase, DeltaLake...), et parfois dans des systèmes développés par l'équipe de l'EDS. La manière dont les clients accèdent à la donnée orientent aussi très fortement ces choix.

Pour l'indexation, un travail similaire est réalisé. Nous utilisons principalement Apache SolR qui fonctionne avec les technologies déjà en place, en particulier HDFS.