DIGORA

Ebook

IoT : Quelles technologies Big Data utiliser ? Pour quels résultats ?

La valeur de l'IoT repose sur les précieuses données générées. Encore faut-il avoir construit une infrastructure Big Data efficace.

Edité le 06/05/20

Table des matières

Comme la téléphonie mobile en son temps, l'IoT tisse peu à peu sa toile dans le quotidien, à la fois des entreprises et des consommateurs.

Mais là où le déploiement réseau « suffisait », c'est aujourd'hui la question de la production, du stockage et du traitement de volumes colossaux de données constamment produites par les objets connectés qui est soulevée.

Car la valeur de l'IoT repose essentiellement sur les données générées par chaque objet connecté. Des données qui sont sources d'informations précieuses, jusqu'à devenir déterminantes dans le cadre de ses activités. Et ces objets connectés sont en pleine expansion : selon IDC et Gartner, ils seront au nombre de 30 milliards en 2020, pour un marché estimé à 1 700 milliards de dollars.

Derrière le réseau d'objets connectés, c'est donc toute une infrastructure de data management, et plus précisément de Big Data, à mettre en place.

Quelles technologies utiliser ? Pour quels résultats attendus ? Couche par couche, voici les recommandations de Digora, en matière de Big Data au service de l'IoT.

Le stockage avec un Cluster Hadoop

Dans cette première partie, parlons des systèmes de stockage pour le Big Data : le cluster Hadoop, associé au système de stockage Blob Azure pour les données froides.

Ce n'est qu'en la combinant à d'autres données que l'on peut maximiser la valeur d'une donnée. C'est justement tout l'objectif d'un cluster Hadoop, qui permet le stockage et l'exploitation de données diverses en grande quantité. Ses multiples outils disponibles (stockage, analyse et traitement : calcul distribué, haute disponibilité) offrent aux entreprises la possibilité de renforcer la valeur de leurs données.

Pour aller plus loin, il est possible de l'associer à du stockage blob (binary large object) tel que le propose Microsoft avec son offre de stockage Blob Azure, conçu pour stocker dans le cloud de très grandes quantités de données non structurées. Dans ce genre d'architectures, Hadoop est essentiellement réservé à la redondance et au calcul distribué (HDFS : Hadoop Distributed File System), tandis que Blob Azure assure l'archivage des données froides.

D'origine Open Source (Apache Software Foundation), il existe de nombreuses distributions Hadoop. Parmi elles, Cloudera a développé, avec Red Hat et Eurotech, une architecture IoT ouverte, complète et modulaire, pour simplifier et accélérer les déploiements IoT. Sécurisée et scalable, cette plateforme, qui embarque les technologies les plus récentes, bénéficie de toute l'expertise d'un éditeur spécialiste du Cloud, tout en restant ouverte pour ne pas risquer de « s'enfermer » dans une technologie propriétaire.

Les bases de données NoSQL : la saine concurrence

En fonction de leurs besoins, les entreprises ont tendance à choisir l'un ou l'autre. Pourtant, le framework Hadoop et la technologie de gestion de bases de données NoSQL sont, malgré les apparences, plutôt complémentaires.

Notamment quant aux workloads qu'ils adressent. En effet, là où la technologie Not Only SQL assure un accès en lecture et écriture ultra rapide favorable aux applications temps réel et aux interactions utilisateurs, Hadoop et sa technologie MapReduce traitent beaucoup plus de données, au détriment des performances.

Parmi les (nombreux) acteurs des bases de données NoSQL, Cassandra, HBase, MongoDB et Couchbase sont les plus connus. Si leurs approches sont globalement similaires, chaque technologie propose des fonctionnalités et des spécificités propres.

- Apache Cassandra : avec son modèle réparti, Cassandra mise prioritairement sur la scalabilité de l'architecture, mais aussi et surtout sur sa haute disponibilité et ses performances de premier plan.

- Apache HBase : conçue d'après le modèle BigTable de Google et développée (en java) dans la cadre du projet Apache Hadoop, HBase est une base de données Open Source, non relationnelle et distribuée. HBase apporte notamment à Hadoop ses capacités de stockage à grande échelle et sa faible latence.

Pour en savoir plus :

Qu'est-ce qu'une base NoSQL ? Les cas Datastax (Cassandra) et MongoDB

Les bases de données Time Series : l'horodatage des données, au service de l'IoT

En matière d'IoT, les bases de données Time Series (ou BDD orientées séries temporelles) trouvent également une grande utilité.

Elles permettent en effet de stocker, visualiser et interroger de grandes quantités de données de séries chronologiques, à l'image de celles générées par les objets connectés.

- InfluxDB : le SGBD InfluxDB propose une architecture distribuée sur plusieurs nœuds, avec des metadata pour la structure. InfluxDB s'insère de façon plus large dans une suite complète baptisée Tick Stack, comprenant également Telegraf (collecte et préparation des données issues des capteurs, terminaux et applications) et Chronograf (dashboards et dataviz).

- Azure Time Series Insights : conçu pour s'intégrer nativement aux passerelles cloud telles que Azure IoT Hub et Azure Event Hubs, Azure Time Series insights gère le stockage et l'interrogation rapide de données. Son explorateur, pour la visualisation de données, et ses API, pour l'intégration des données à des applications personnalisées, en font l'un des incontournables des BDD time series du marché.

Les Message-Oriented Middlewares (MOM) ou brokers de messages

Afin d'éviter de connecter un par un chaque système émetteur de données à chaque système consommateur de données, les MOM (Message-Oriented Middlewares) prennent la forme de bus, auprès desquels les émetteurs publient leurs messages, et sur lesquels les consommateurs viennent les lire.

À noter qu'aucun traitement n'est réalisé dans ce processus : le principal intérêt de ce dispositif est l'asynchronisme, ne nécessitant pas la disponibilité de chaque partie pour assurer l'échange (le rôle du MOM étant de conserver le message jusqu'à la sa transmission au destinataire).

Si ActiveMQ ou Joram peuvent suffire dans de nombreux cas, 3 solutions plus avancées permettent de répondre à des besoins plus spécifiques.

- RabbitMQ : soutenu par Pivotal et utilisé notamment par Instagram et Clever Cloud, RabbitMQ adresse en particulier les processus publish/subscribe traditionnel ou nécessitant un routage élaboré des messages. Et peut gérer en outre l'utilisation des protocoles spécifiques tels que STOMP (Streaming Text Oriented Messaging Protocol), idéal pour le temps réel, et MQTT (Message Queue Telemetry Transport), garantissant une utilisation minimale de la bande passante, parfait pour l'IoT.

- KAFKA : créé (et utilisé) par LinkedIn, KAFKA est également à l'œuvre chez Uber, Netflix, AirBnB, OVH ou encore Goldman Sachs. Capable d'ingérer une grande quantité de données en un laps de temps minimum, de conserver les messages plus longtemps et de capter des événements (tels qu'un changement dans une base de données par exemple), KAFKA est également en mesure de répondre à des besoins transactionnels.

  De la même façon, Oracle Golden Gate et IBM IDR, qui sont deux technologies de Change Data Capture (CDC) peuvent compléter KAFKA pour le suivi de changements en bases de données.

- Azure IoT Hub : capable de prendre quasiment tout type d'objets connectés, Azure IoT Hub est un concentrateur de messages proposant une communication bidirectionnelle entre une application et les objets auxquels elle est connectée. IoT Hub est un service managé, hébergé dans le cloud.

Comme évoqué précédemment, le protocole MQTT est particulièrement adapté à l'internet des objets, car les messages qu'ils transportent sont particulièrement légers (256 Mo max.). Dans une logique publish/subscribe, les clients abonnés ont le choix entre plusieurs niveaux de qualité de service (suivi des messages et garantie de remise au destinataire).

Le Machine Learning pour valoriser les données

Utiliser l'IoT pour capter et stocker des données ne suffit pas : reste ensuite à les valoriser, en particulier par la création de nouvelles données et informations résultant de leur analyse.

À cette fin, les outils de Machine Learning sont aujourd'hui indispensables à l'analyse tant les volumes sont importants, ce qui rend leur gestion inaccessible aux humains. Dans tous les cas, la première question à se poser est l'usage : doit-on utiliser du Machine Learning « classique » par apprentissage statistique, ou bien du Deep Learning sur fond de réseaux de neurones ?

Quant aux langages utilisés, si de nombreuses initiatives ont vu le jour, les cas d'usages les plus avancés ont été réalisés en Python, qui se démarque clairement aujourd'hui. Pas moins d'une quinzaine de frameworks s'appuient sur ce langage. Parmi lesquels Caffe et Tensorflow pour les plus connus ou encore Azure ML Service et Apache Spark :

- Azure ML Service : le service machine learning d'Azure embarque 2 composants. ML Studio en premier lieu, dont l'interface graphique permet de tester rapidement différents modèles. Azure Databricks de son côté, basé sur Apache Spark (voir ci-dessous) est utilisé pour le « nettoyage » de données (data engineering) et le machine learning, dans des langages tels que Python et R.

- Apache Spark : Utilisable avec Python pour l'écriture des tâches, Apache Spark reste toutefois mieux adapté au langage Scala et intégré à Hadoop. On notera l'intégration de TensorFlow à Apache Spark

L'acheminement et le traitement des données

Par définition, l'IoT implique l'utilisation de plusieurs systèmes : l'objet connecté en lui-même d'une part, et tous les systèmes pouvant utiliser les données qu'il émet, pour les collecter, les stocker et les traiter.

Dès lors, la construction de solutions s'appuie sur des workflows de données, qu'il s'agit de construire. C'est tout l'objet des « workflows tools », parmi lesquels Node-RED et Apache Nifi.

- Node-RED : basé sur Node.js, Node-RED est une application de conception de chaînes de traitement en environnement Web, principalement destiné aux Proof Of Concept ou aux architectures de faible envergure. Au même titre qu'un ETL, Node-RED embarque une palette de connecteurs, des composants de traitement, et des modules pour lier entre eux des terminaux physiques, des API et des services en ligne.

  La force de Node-RED : la possibilité d'installer de nouveaux composants depuis la liste officielle des composants Node-RED, ou d'en développer soi-même pour répondre à ses besoins spécifiques. La création (en drag & drop) et le déploiement (en un clic) d'un « flow » s'opèrent entièrement depuis l'interface graphique.

- Apache Nifi : Avec son projet Nifi, soutenu par Hortonworks, la fondation Apache prouve, s'il en était besoin, son implication dans le développement des dispositifs IoT. Concrètement, Nifi permet d'injecter automatiquement des flux de données entre différents systèmes sources en direction d'autres systèmes cibles.

  Capable de prendre en charge l'intégralité des flux de données, très tolérant aux pannes et scalable pour la gestion de très grands volumes, Nifi dispose d'une interface web complète pour la définition et le contrôle en temps de l'acheminement des données.

Identifiez la valeur de votre projet IoT/Big Data

Les projets IoT et Big Data portent avec eux de nombreux enjeux business et technologiques, comme nous venons de le voir. Afin de préparer et développer au mieux votre projet, il s'agit surtout d'identifier la valeur que pourra générer ce projet pour votre entreprise, vos métiers et/ou vos clients.

Nos experts Digora vous accompagnent pour réaliser une étude d'opportunité de votre projet Iot/Big Data, grâce à une méthodologie de Design centrée utilisateur afin de :

- Diagnostiquer et Comprendre vos métiers et votre stratégie,
- Sensibiliser à l'IoT et lancer un processus d'idéation des cas d'usages,
- Prioriser et se projeter sur les cas d'usages les plus intéressants et rentable pour vous.

Cette méthode s'articule autour d'ateliers qui se déroulent chez vous et avec vous pour :

- Définir le périmètre,
- Définir la valeur,
- Identifier la valeur,
- Auditer les SI,
- Restituer la synthèse des projets à travers un business plan définissant le ROI, RONI ou ROI Inveré.

→ Je contacte un expert Digora

Digora, Experts de la gestion des données

Reconnu comme un expert de l'administration et l'optimisation des bases de données, Digora accompagne les entreprises sur leurs enjeux de performance, de sécurité, et de disponibilité du SI, tout en intégrant les technologies Cloud et de l'IoT.

Digora propose des services de Conseil & Transformation Numérique, de fourniture et d'hébergement d'infrastructure, de Services Managés (Support Technologique, Maintien en Condition Opérationnelle) et d'Innovation Digitale avec une plateforme IoT Hub.

Créée en 1997 et ayant son siège à Strasbourg, Digora est présente en France (Bordeaux, Lille, Lyon, Paris, Rennes, Strasbourg et Toulouse), au Luxembourg et au Maroc.

Fiers de ses 140 collaborateurs, Digora est une entreprise à taille humaine et innovante dont les collaborateurs ont à cœur d'accompagner leurs clients dans la réussite de leur transformation digitale.

En 2019, Digora a réalisé un chiffre d'affaire de près de 23 millions d'euros.

Digora compte 550 clients actifs, grands comptes et ETI de tous secteurs d'activités dont : BANDAI NAMCO, BNP Paribas, Compagnie des Alpes, le Conseil Régional d'Aquitaine, Engie, le Groupe ÏDKIDS, Lacoste, Ramsay - Générale de Santé, Maincare, Maisons du Monde, XPO, Poclain Hydraulics, Sanofi, Toyota, l'UGAP. Découvre quelques beaux projets dans la rubrique Presse.

Digora a également noué des partenariats solides avec des acteurs forts de l'IT Une équipe à la pointe, Ce qui fait de Digora un expert des infrastructures IT et de la gestion de données.

→ Je contacte un expert Digora