Le Data Warehouse fait sa révolution digitale

Depuis longtemps déjà, les grandes entreprises ont installé des data warehouse pour stocker les données de leurs clients. Mais la révolution digitale est passée par là et les appliances spécialisées font place à des infrastructures 100 % logicielles et 100 % Cloud.

La data, c'est le pétrole de la révolution digitale. La formule est bien connue, mais c'est une réalité. Depuis une dizaine d'années, toutes les grandes entreprises cherchent à exploiter les données pour mieux connaître leurs clients et pour optimiser leurs opérations.

Néanmoins, ce que l'on nomme data dépasse très largement le cadre des données structurées qui constituaient le fondement des systèmes décisionnels dans les années 1990.

Les entreprises veulent croiser les informations issues de leurs systèmes ERP et CRM, mais aussi des données générées par des objets connectés, des verbatim glanés dans les centres d'appels ou sur les réseaux sociaux, ce que les data warehouse traditionnels étaient bien incapables de faire. Ainsi, de nombreuses entreprises ont bâti un data lake en s'appuyant sur Hadoop, une pile logicielle Open Source qui a donné aux entreprises l'accès aux technologies big data directement issues des laboratoires de Google.

Parmi les principaux acteurs à s'être positionnés sur ce marché et à fournir un support de classe Entreprise sur Hadoop, IBM, mais aussi MapR Technologies, entreprise qui vient d'être rachetée par HPE, et Cloudera qui s'est imposé comme le porte-drapeau de ce marché.
La start-up, cofondée par Jeff Hammerbach, le grand gourou des mathématiques de Facebook, a fusionné avec son plus grand rival, Hortonworks, en octobre 2018.

Denis Fraval, directeur avant-vente pour Cloudera, souligne : « Notre plateforme est 100 % Open Source. C'est ce qui permet aux entreprises de s'affranchir de toute adhérence avec l'infrastructure.

Une société qui a déployé Hadoop sur une infrastructure on-premise et qui souhaite, demain, utiliser les services d'un fournisseur cloud public tels qu'Amazon Web Services, Microsoft Azure ou Google Cloud Platform pourra le faire et ce choix restera totalement réversible. Elle pourra exécuter ses traitements big data au moyen d'une interface unique pour gérer ses environnements on-premise et cloud. » Parmi les grands axes de développement de Cloudera : le support des trois gros acteurs du cloud à isofonctionnalité, le portage de son offre sous forme de conteneurs logiciels pour faciliter un peu plus encore sa portabilité.

Base de données, couleur Big data

Si big data fut un temps synonyme de Hadoop, ce n'est plus le cas aujourd'hui. Tous les grands éditeurs de solutions de data warehouse ont fait significativement évoluer leurs offres pour apporter des solutions qui gèrent les gros volumes de données, mais aussi la diversité des données manipulées par les algorithmes.

Teradata, un acteur historique du data warehouse propose toujours son offre hardware avec les Teradata Intelliflex, des équipements d'architecture MPP particulièrement optimisés pour l'analytique et les hautes performances. Néanmoins, Teradata propose sa base de données sous la forme d'un logiciel pour des déploiements on-premise ou dans le cloud.

Cédric Giacomo, Solutions Engineer chez Teradata France souligne : « Notre offre est disponible sur Microsoft Azure, Amazon Web Services et le support de Google Cloud arrive en 2020. L'objectif est d'offrir un maximum de flexibilité dans le choix de solution. Un client peut ainsi démarrer un projet sur un mode on-premise, puis basculer dans le cloud sans aucune difficulté, avec les mêmes conditions financières. » L'éditeur évolue vers un modèle de facturation à la consommation, quelle que soit l'infrastructure.

Disposer du même logiciel côté on-premise et côté cloud permet la sauvegarde/récupération des données d'un côté vers l'autre. Il permet encore de travailler sur une architecture réellement hybride et d'interroger des données hébergées par un système Teradata dans le cloud.

Parmi les différenciants de la base de données Teradata, la maîtrise des traitements parallèles qui reste, selon Clément Droinat, Solution Engineer de Teradata, un atout majeur : « Notre workload management est bien plus efficient que celui de nos concurrents, ce qui permet de traiter davantage de requêtes en simultané, avec une puissance donnée. » Enfin, si la base de données Teradata est, avant tout, conçue pour porter des données structurées, l'éditeur met à disposition des composants pour accéder à des données stockées sur le service de stockage Amazon S3, dans des clusters Hadoop.

Les principales solutions de data management pour l'analytique selon Gartner

Éditeur	Solutions
Alibaba Cloud	ApsaraDB RDS for MySQL, SQL Server and PostgreSQL. HybridDB for PostgreSQL (base sur la base de données Greenplum) MaxCompute pour les grands data warehouse. E-MapReduce for Hadoop.
Amazon Web Services	Amazon RedShift, service de data warehouse. Amazon Elastic MapReduce (EMR), service Hadoop managé.
Arm Treasure Data	Customer Data Platform (CDP). Solution de data management portée par AWS.
Cloudera	Suite logicielle Cloudera Data Platform, une distribution Hadoop.
GBase	L'éditeur chinois propose GBase 8a, une base relationnelle MPP dédiée au data warehouse. GBase HD, une distribution Hadoop. GBase UP, plateforme de data warehouse virtuel (LDW).
Google	BigQuery : offre de data warehouse serverless. Cloud dataproc : services Hadoop et Spark managés. Cloud dataflow sur le traitement batch ou en streaming des données.
Huawei	Plateforme FusionInsight, intégrant une suite Hadoop. GaussDB 200 une base de données MPP propriétaire.
IBM	Outre sa base de données DB2, IBM propose des appliances PureData et des systèmes intégrés pour l'analytique. L'américain propose aussi des solutions Hadoop BigInsights et DB2 Warehouse sur son cloud.
MapR Technologies (HPE)	Suite logicielle Hadoop Converged Data Platform (CDP).
MarkLogic	Base de données disponible en édition développeur gratuite ou édition entreprise pouvant être déployée indifféremment on-premise ou dans le cloud.
Micro Focus	Base de données en colonne Vertica Enterprise disponible sous forme logicielle classique ou sous forme de service cloud.
Microsoft	Outre sa base de données SQL Server, Microsoft propose le service Microsoft Azure Synapse Analytics, ainsi qu'Azure data lake et Azure HDInsight, distribution managée par Microsoft et basée sur Hortonworks.
Neo4J	Base de données orientée graphe, éditée en Open Source, qui peut s'interfacer avec Apache Spark, via son module CAPS.
Oracle	Outre sa base de données Oracle 19c et appliances Exadata et Big data Appliance, Oracle pousse son offre Oracle Autonomous Data Warehouse (ADW) cloud.
Pivotal	Base de données MPP Greenplum basée sur PostgreSQL.
SAP	L'éditeur propose HANA, une base de données en colonnes au fonctionnement in-Memory. SAP BW/4HANA constitue l'offre dédié au data warehouse.

Oracle mise sur le Machine Learning

Autre acteur incontournable du marché de la base de données, Oracle mise sur les capacités d'auto-administration de sa base de données pour se distinguer de ses concurrents.

« Notre stratégie consiste à apporter des réponses à des besoins métier et donner le pouvoir aux directions fonctionnelles. L'équipe RH d'une grande organisation aura à sa disposition une application pour gérer les données sur ses Top Talents, et une autre application avec les données de paye. Ces équipes doivent s'appuyer sur leurs équivalents IT pour faire communiquer ces deux silos. Avec ce que nous leur proposons, nous allons leur donner la main, afin de créer un datamart qui permette de mener leurs analyses et prendre leurs décisions », explique Karim Zein, vice-président d'Oracle France.

Cette proposition de valeur s'appuie sur la technologie de machine learning mise en œuvre dans Autonomous Data Warehouse (ADW), une IA qui permet de s'affranchir de toutes compétences techniques pour créer et paramétrer des datamarts, anticiper les besoins en capacités de calcul ou de stockage, mettre en place des mesures de sécurité pour protéger les données et, enfin, assurer la haute disponibilité de cette infrastructure de données.

En outre, Oracle pousse désormais ses offres cloud en priorité vis-à-vis de ces utilisateurs métiers : « Nous proposons cette offre en priorité sur notre cloud public et ADW se consomme comme une application SaaS, pour être le plus accessible possible à des non-experts. » Une version similaire d'ADW arrive sur une offre matérielle destinée à être exploitée dans le data center des entreprises, une offre qui va rejoindre la gamme Cloud at Customer d'Oracle.

Le Cloud va-t-il tout emporter ?

Les éditeurs traditionnels mettent les bouchées doubles pour moderniser leurs offres, mais un mouvement de fond agite aujourd'hui le secteur. De nombreuses entreprises préfèrent créer leur data lake sur le cloud public. Elles profitent ainsi des capacités de stockage et de montée en puissance, qui restent très coûteuses à obtenir dans un data center privé .

Avec son offre Amazon EMR, AWS propose une plateforme big data complète intégrant Apache Hadoop et de nombreux composants Open Source. Euronext exploite, de son côté, un data lake Amazon ERM pour gérer ses données de trading.

Le leader mondial du cloud propose, par ailleurs, Redshift, une solution de data warehousing qui bénéficie à plein de l'écosystème de solutions cloud d'AWS, puisqu'il peut être exploité en lien avec Amazon ERM, l'outil de requêtage Amazon Athena, et encore les outils dédiés à l'intelligence artificielle Amazon SageMaker.

Enfin, pour les entreprises qui souhaitent rapidement construire un data lake en partant d'une page blanche, AWS Lake Formation est un outil de conception rapide de lac de données. Une logique que Microsoft suit exactement. En effet, l'éditeur a porté SQL Serveur pour data warehouse sur Azure et commercialise l'offre sous le nom d'Azure Synapse Analytics, qui cohabite avec Azure Data Lake et Azure HDInsight, une distribution Hadoop managée par Microsoft. Là encore, la

synergie avec l'ensemble des outils de traitement et d'analyse des données présents sur Azure crée de la valeur autour de l'offre de stockage de données proprement dite.

De son côté, Google propose la solution BigQuery sur son cloud. « BigQuery permet l'analyse de quantités massives de données en abstrayant complètement l'infrastructure sous-jacente. En outre, la solution offre une interface SQL familière et compatible avec les outils du marché », argumente Bastien Legras, directeur technique & customer engineer lead chez Google Cloud France.

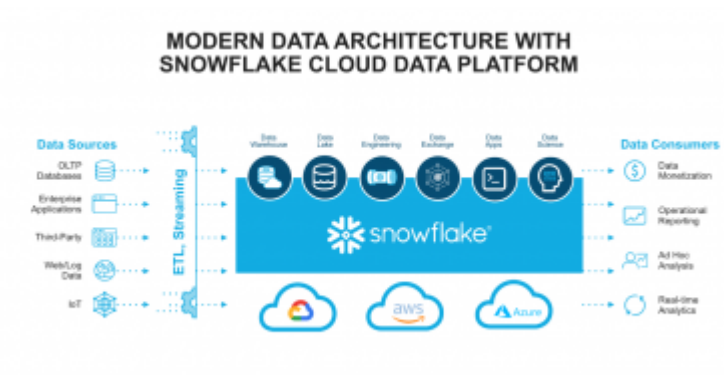
Par ailleurs, l'offre est complétée, entre autres, par Dataproc, la solution de modernisation de data lake permettant de capitaliser sur les développements Spark et Hadoop réalisés sur les data lake et de bénéficier de moteurs de traitement de la donnée en temps réel, avec ou sans code, comme Dataflow et Data Fusion.

Autre atout de l'offre, son intégration dans les technologies d'IA de Google : « BigQuery s'intègre particulièrement bien avec nos solutions Cloud AI et, notamment, AutoML Tables qui permet de déployer des modèles de machine learning sur des données structurées sans avoir à coder. On peut développer et, surtout, rendre opérationnels des modèles de machine learning en quelques jours au lieu de plusieurs semaines. »

Nouvelle génération

Un vent d'innovation souffle désormais sur ce marché, à l'image de la base de données 100 % cloud Snowflake conçue à partir de 2012 par d'anciens ingénieurs français d'Oracle pour répondre aux besoins spécifiques de l'analytique.

« Snowflake a été écrite à partir d'une feuille blanche, avec une architecture interne totalement différente de celle des bases de données classiques. Cette architecture assure une séparation complète entre calcul et stockage, une approche conçue dès l'origine pour le Saas et qui ne tourne que dans le cloud », explique Olivier Leduc, son directeur technique pour la France.



La solution est disponible chez les trois principaux fournisseurs de cloud public, la compatibilité avec Google Cloud devant être effective cet été. Elle s'adresse autant à des start-up qui veulent stocker des volumes de données relativement modestes, mais aussi aux grands data ware house des groupes internationaux.

« Notre architecture est conçue pour être totalement élastique et sa subtilité réside dans la mise en œuvre d'un pool de ressources, dans lequel chaque client vient piocher. Cette approche permet à de multiples clients d'utiliser l'infrastructure cloud en bénéficiant de performances très élevées puisqu'il ne faut pas démarrer de nouvelles machines à chaque fois qu'un client sollicite des ressources » poursuit Olivier Leduc.

La solution est conçue pour stocker des données structurées ou semi-structurées de type documents Json, fichiers au format Apache Parquet, par exemple. À noter la fonction de data sharing qui permet à une entreprise utilisant Snowflake de donner à des tiers l'accès à un dataset, un ensemble de données bien définies.

Autres acteurs de la nouvelle vague qui s'intéressent aujourd'hui au monde du data warehouse, MariaDB propose une édition dédiée à l'analytique s'appuyant sur le service de stockage Amazon S3 ou, encore, MongoDB qui, lui aussi, s'intéresse à ce marché des données »froides ». Cette base de données NoSQL est très appréciée, notamment, sur le Web : c'est elle qui est »derrière » le site marchand de Leroy Merlin, mais également dans les infrastructures de Facebook ou d'eBay.

47 zettaoctets à stocker en 2020

MongoDB a récemment dévoilé Atlas Data Lake, une solution dédiée aux données froides et aussi un data lake que les développeurs MongoDB vont pouvoir requêter exactement de la même manière que sur la base de données de production.

« La solution va utiliser des données stockées sur Amazon S3 sans avoir à les déplacer. De multiples formats sont supportés, de même que cela va fonctionner sur des données compressées » explique Julien Contarin, senior solutions architect chez MongoDB France, « les mêmes requêtes peuvent s'exécuter sur S3 et sur MongoDB. Cette solution est ainsi compatible avec l'ensemble de l'écosystème MongoDB. »

La solution est actuellement en bêta sur Amazon Web Services et sera bientôt portée sur Microsoft Azure et Google Cloud Platform. Marché né il y a près de 40 ans, le data warehouse ne cesse de se réinventer, ce qui n'a rien d'un luxe lorsqu'on sait que l'humanité va devoir stocker 47 zettaoctets en 2020… et 175 Zo en 2025 !