



# Qubole Data Services

BIG DATA | Décembre 2021

**1** Aujourd'hui, le monde où nous vivons actuellement, nous créons des données massivement et quotidiennement. Plusieurs problèmes surviennent alors : comment stocker toutes ces données?, Comment avoir une vision globale sur celle-ci?, et comment en extraire et traiter les données? C'est pour cela que le Big Data existe pour répondre à toutes ces problématiques à l'aide de ses nombreux outils !

**2** Une meilleure connaissance de ses clients, des campagnes marketing plus efficaces, une expérience utilisateur personnalisée, une meilleure fidélisation, la prédiction des tendances. Tous sont des avantages pour les entreprises prenant en compte les apports du Big Data sur ses concurrents. Il est donc important de nos jours pour qu'une entreprise fonctionne d'étudier le Big Data.

**3** Pour tout type d'études du Big Data, différents outils sont mis en place et dans cet article nous allons nous intéresser à un nouvel outil de gestion du Big Data, utilisant des fonctionnalités Open Source, qui s'appelle **Qubole**

## THE NEW OPEN DATA LAKE PLATFORM

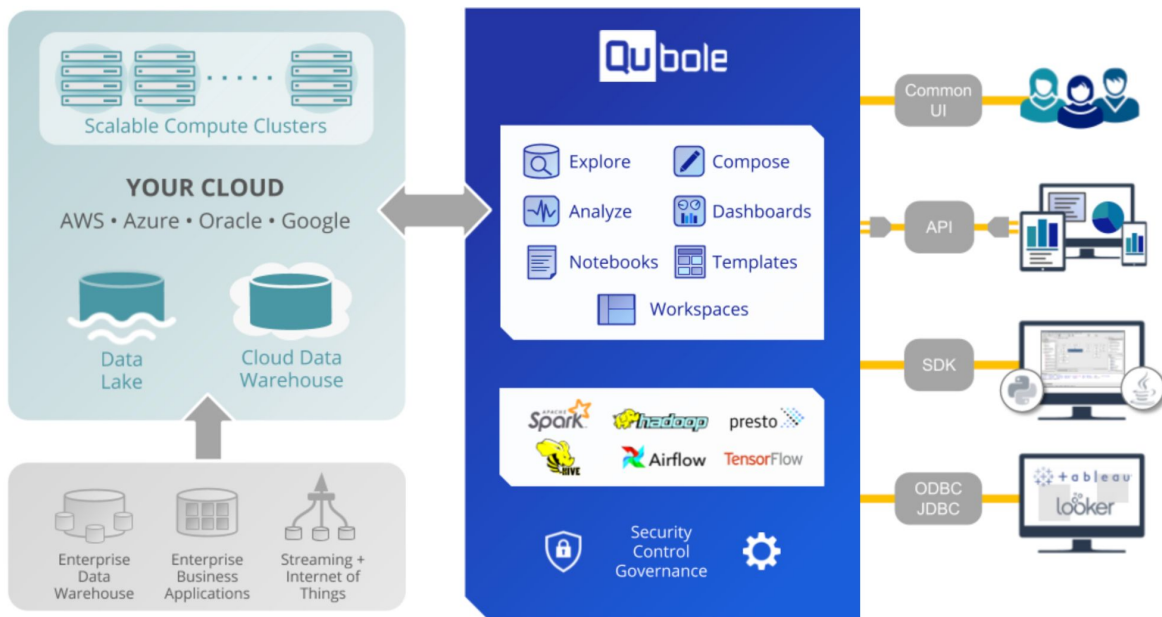
Qubole se présente comme "Une plate-forme data lake simple, sécurisée et open source pour l'apprentissage automatique (Machine learning), le streaming et l'analyse ad-hoc" par ses développeurs. Un data lake est défini comme ~~est~~ un système ou un référentiel qui stocke des données dans son format brut ainsi que des ensembles de données de confiance transformés et fournit un accès à la fois programmatique et basé sur SQL à ces données pour diverses tâches d'analyse telles que le Machine Learning, l'exploration de données et l'analyse interactive.

# APPORTS & AVANTAGES

Ce qu'apporte principalement Qubole c'est avant tout la flexibilité de la charge de travail des données tout en accélérant radicalement l'adoption des data lake, en réduisant le délai de rentabilisation et en réduisant les coûts des data lake dans le cloud de 50 %. La plate-forme de Qubole fournit des services des data lake de bout en bout tels que la gestion de l'infrastructure cloud, la gestion des données, l'ingénierie continue des données, l'analyse et l'apprentissage automatique avec une administration quasi nulle. Qubole a la confiance de grandes marques telles qu'Expedia, Disney, Oracle et Adobe pour stimuler l'innovation et transformer leurs entreprises à l'ère du big data.

La plate-forme cloud de Qubole s'exécute sur tous types d'infrastructures cloud, en tirant parti d'un cloud privé virtuel (AWS ou Google), d'un réseau virtuel (Microsoft Azure) ou d'un réseau cloud virtuel (Oracle).

Vous trouverez ci dessous un schéma expliquant quels sont les utilités de la plateforme que propose Qubole. On peut voir que l'application proposent pas mal de features comme: un notebook, une analyse des données ou encore des Dashboards



## VALEURS

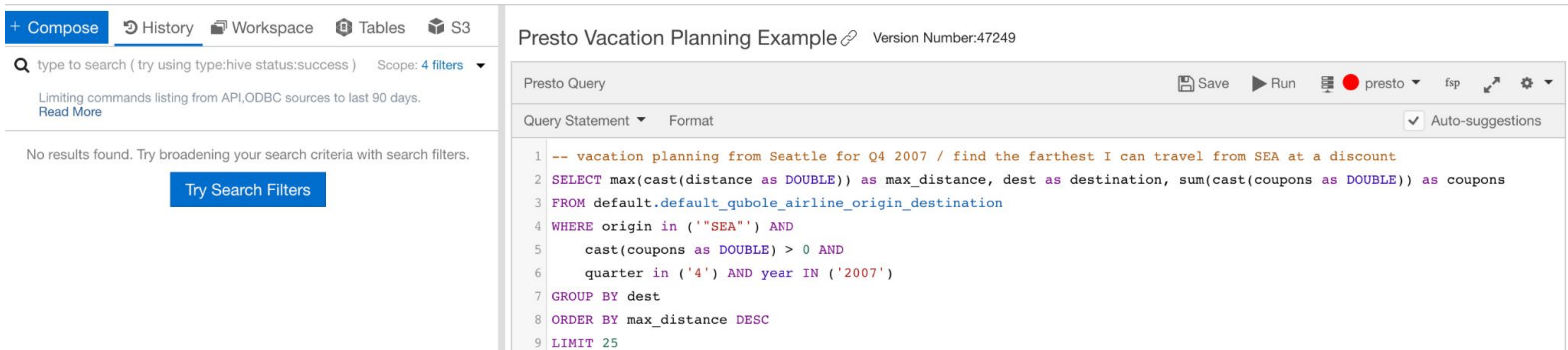
Qubole est une société qui se veut à l'écoute de ses clients et est composé d'une équipe expérimentée travaillant tous dans l'univers du big data. Gagner du temps et la valeur principale de Qubole car "le temps est précieux". Qubole permet aussi de travailler sur des projets très constructif en ayant de l'impact sur les résultats proposés. Mais Qubole c'est aussi anticiper des faits et avoir de l'avance sur ce que vous faites !



# QUBOLE FEATURES

Nous avons tenté d'explorer au mieux toutes les features que Qubole met en avant et expliquer au mieux notre ressenti

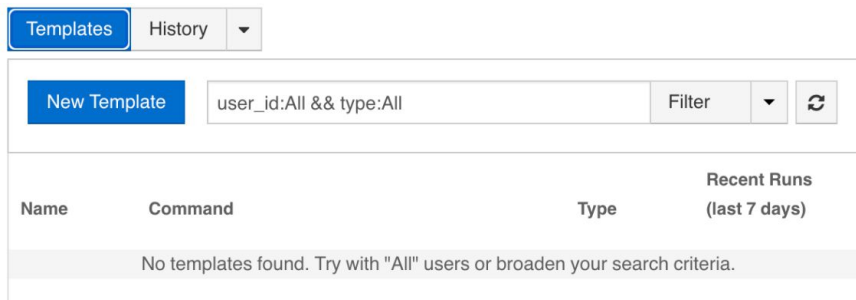
Nous commençons avec la feature "Compose", celle-ci permet d'afficher un éditeur de requête très pratique pour exécuter une grande variété de commande comme pour utiliser des commandes shell ou encore interroger des frameworks open source tel Apache Hive, Hadoop ou Spark.



The screenshot shows the Qubole Compose interface. At the top, there are tabs for '+ Compose', 'History', 'Workspace', 'Tables', and 'S3'. Below these, a search bar is visible with the text 'type to search ( try using type:hive status:success )' and a scope filter set to '4 filters'. The main area is titled 'Presto Vacation Planning Example' with a version number of 47249. It contains a Presto query editor with a query statement that selects the maximum distance from Seattle for Q4 2007, based on coupons and destination. The query is as follows:

```
1 -- vacation planning from Seattle for Q4 2007 / find the farthest I can travel from SEA at a discount
2 SELECT max(cast(distance as DOUBLE)) as max_distance, dest as destination, sum(cast(coupons as DOUBLE)) as coupons
3 FROM default.default_qubole_airline_origin_destination
4 WHERE origin in ('SEA') AND
5       cast(coupons as DOUBLE) > 0 AND
6       quarter in ('4') AND year IN ('2007')
7 GROUP BY dest
8 ORDER BY max_distance DESC
9 LIMIT 25
```

Nous avons ensuite la feature "template" où sont répertorié plusieurs modèles de commandes Qubole. Elle permettrons de pouvoir composer une commande ou une requête plus facilement et en une seule fois. On pourra ensuite modifier cette commande autant de fois que l'on veut en changeant les paramètres. Mais lors de notre essaie nous avons eu quelques difficultés avec Qubole pour créer de nouveau Template le système n' étant pas assez fiable



The screenshot shows the Qubole Templates interface. It has tabs for 'Templates' and 'History'. Below the tabs, there is a 'New Template' button and a search bar containing 'user\_id:All && type:All'. To the right of the search bar are 'Filter' and 'Refresh' buttons. Below the search bar, there is a table with columns 'Name', 'Command', 'Type', and 'Recent Runs (last 7 days)'. The table is currently empty, with a message 'No templates found. Try with "All" users or broaden your search criteria.'

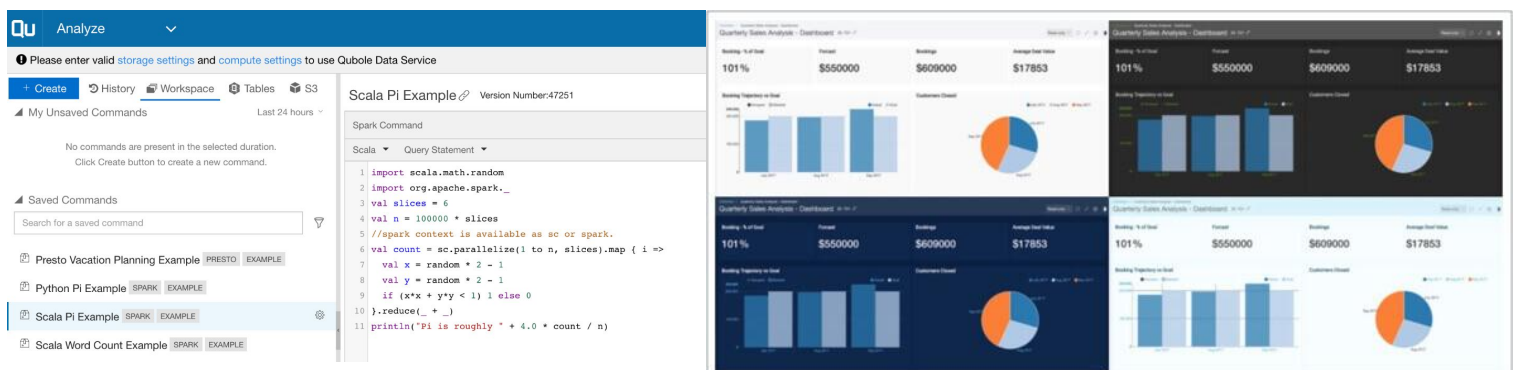
We're sorry, but something went wrong.

We hit an error on one of our servers.

The Qubole application experienced an issue in processing this request and it was not successful. If you would like this issue investigated, please provide error code and precise details in a Qubole support ticket for further analysis or send a mail via Report Issue link given below.

[Report Issue](#)

Nous avons essayé ensuite la feature "Analyze" qui permet de composer et exécuter des requêtes ou des commandes sur des ensembles de données. Il est également utilisé pour échanger des données entre différentes bases de données à l'aide des commandes d'importation ou d'exportation de données. Nous pouvons ensuite grâce à ces données créer des Dashboards qui permettront une analyse plus visuel ce qui est vraiment une plus valus pour ce type de plateforme. On peut ensuite incorporer de la technologie machine learning.



The screenshot shows the Qubole Analyze interface. It has a top bar with 'Qu' and 'Analyze'. Below this, there is a message 'Please enter valid storage settings and compute settings to use Qubole Data Service'. The main area is titled 'Scala Pi Example' with a version number of 47251. It contains a Scala command editor with a command that calculates the value of Pi using a random number generator. The command is as follows:

```
1 import scala.math.random
2 import org.apache.spark._
3 val slices = 6
4 val n = 100000 * slices
5 //spark context is available as sc or spark.
6 val count = sc.parallelize(1 to n, slices).map { i =>
7   val x = random * 2 - 1
8   val y = random * 2 - 1
9   if (x*x + y*y < 1) 1 else 0
10 }.reduce(_ + _)
11 println("Pi is roughly " + 4.0 * count / n)
```

Below the command editor, there are four dashboards showing the results of the command. Each dashboard displays a bar chart, a pie chart, and a table of results. The results are as follows:

Running % of total	Processed	Estimated	Average Task Value
101%	\$550000	\$609000	\$17853





La Plateforme Qubole est vraiment une technologie très innovante est prometteuse. Les features proposés sont vraiment intéressantes dans l'univers du Big Data et fonctionnent plutôt bien malgré quelques bugs.

Dans cet article nous n'avons pas pu tester l'intégralité des features comme celle du Notebook qui permet de réexécuter des requêtes, des rapports et des visualisations interactives à l'aide de Zeppelin ou de Jupyter

Des mises à jour sont faites assez régulièrement avec de nouvelles features disponible en bêta comme la possibilité de créer des pipelines sous git.

## CONCLUSION ET AVIS QUBOLE



La Plateforme Data Lake qui est Qubole fait très bien la bascule entre le cloud et les API, Common UI, SDK et ODBC/JDBC utilisé à l'aide de nombreuses features intéressantes. Le tout avec une interface très fluide. Qubole est une bonne plateforme pour exprimer librement son esprit créatif sur projet. Il faut avoir à l'avenir comment se développe l'application et quels sont les nouvelles features intégrés ce qui pourrait faire de Qubole un excellent outil à utiliser en Big Data.