Formation Apache Doris

De la théorie à la pratique

Présenté par : Eugénie Barlet, Aya Mecheri, Nico Dena et Perrine Ibouroi

22 octobre 2025

Déroulement de la séance

1 **Présentation**
Introduction à Apache Doris et concepts clés
~15 minutes

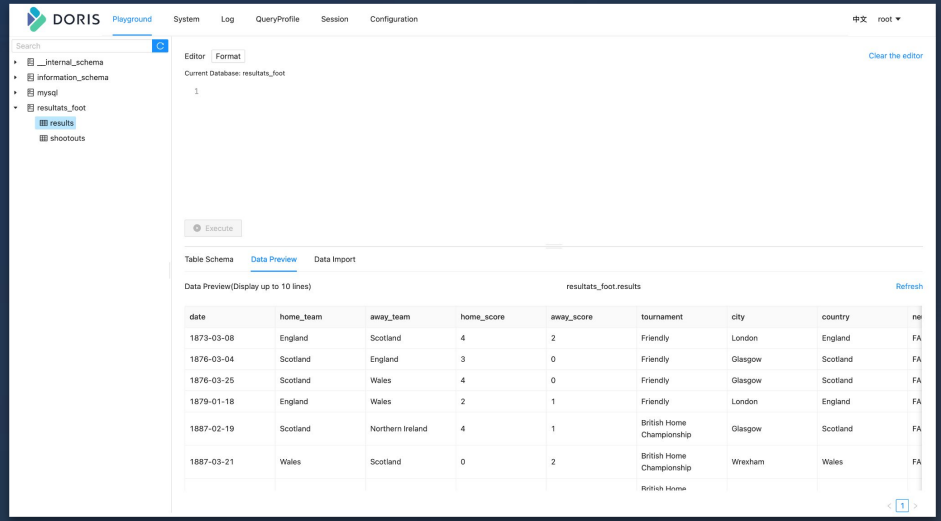
2 **Démonstration**
Installation et première prise en main
~15 minutes

3 **Travaux Dirigés**
Analyses OLAP les résultats des matchs internationaux de football
~120 minutes

4 **Correction**
Débriefing et mise à disposition du corrigé

Qu'est-ce que Apache Doris ?

Un entrepôt de données **OLAP** en temps réel basé sur l'architecture **MPP** (Massive Parallel Processing), connu pour sa rapidité d'exécution des requêtes.



⚡ **Requêtes ultra-rapides**
Sub-secondes - Scannez 1 milliard de lignes en < 1 seconde

👥 **Haute concurrence**
Des milliers d'utilisateurs simultanés sans dégradation

📊 **Analyses complexes à haut débit**
Jointures, agrégations, window functions en temps réel

5000+
Entreprises utilisatrices

600+
Contributeurs

120+
Contributeurs actifs/mois

🚀 **3-10x plus rapide** que les solutions OLAP traditionnelles

🔧 **Installation simplifiée** vs ClickHouse

💾 **Compatible MySQL** Protocole et syntaxe

De projet interne à leader mondial

🚀 **2008-2013**

Origines chez Baidu

Créé sous le nom de **Palo** pour gérer le reporting publicitaire de Baidu.

Développé en interne pour traiter **des milliards de requêtes quotidiennes** en temps réel.

🌐 **2017**

Open Source

Officiellement open-sourcé par Baidu sur **GitHub**, rendant la technologie accessible à la communauté mondiale.

🏛️ **Juillet 2018**

Apache Incubator

Rejoint l'Apache Software Foundation en tant que **projet incubateur**. Renommé de "Palo" à "Apache Doris" et développé sous la supervision de mentors Apache.

🎓 **Juin 2022**

Top-Level Project

Diplômé avec succès de l'incubateur Apache et devient un **projet de premier niveau** (Top-Level Project).

🌟 **Aujourd'hui (2025)**

Écosystème Mature

Utilisé en production par **5000+ entreprises** dans le monde entier, incluant des géants comme TikTok, Xiaomi, Tencent, et Meituan. Adoption massive dans la finance, le retail, les télécoms, l'énergie et la santé.

Apache Doris dans l'écosystème data moderne

Le parcours de la donnée

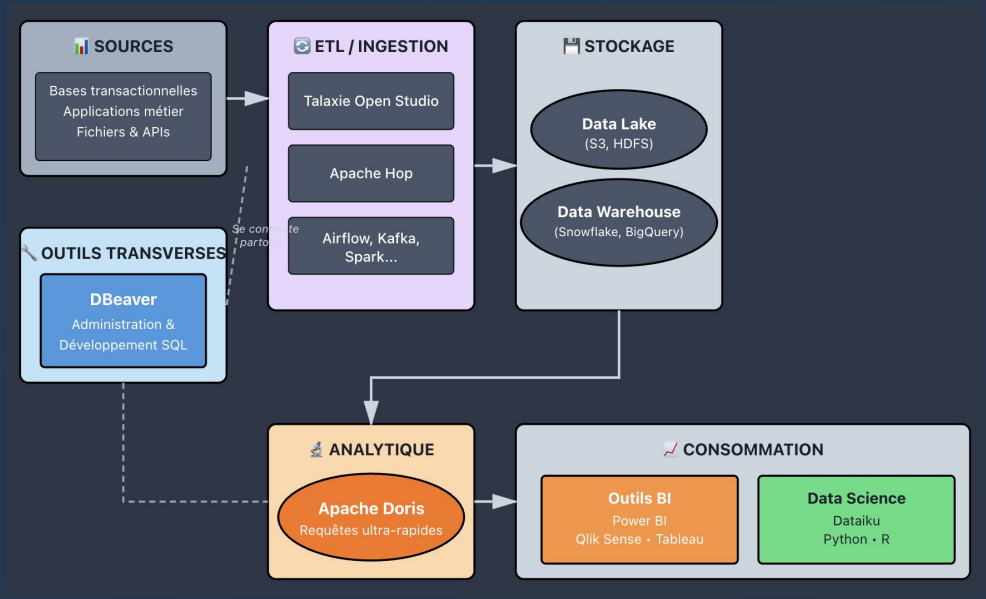
Sources → Données brutes
Bases transactionnelles, APIs, fichiers, applications métier

ETL → Extraction et transformation
Extraction et transformation des données via des pipelines
- **Talaxie Open Studio** : orchestration visuelle des flux
- **Apache Hop** : pipelines de données

Stockage → Persistance
Stockage des données selon leur nature
- **Data Lake** : données brutes et non structurées
- **Data Warehouse** : données structurées

Apache Doris → Analyse ultra-rapide
Le moteur de requêtes temps réel qui lit directement depuis le stockage et répond en sub-seconde aux questions analytiques complexes.

Consommation → Utilisateurs finaux
Restitution via les outils métier
- **BI** (Power BI, Qlik) : dashboards et rapports
- **Data Science** (Dataiku) : ML et analyses avancées



Une architecture minimaliste et puissante

> Apache Doris possède une **architecture minimaliste** avec seulement **deux types de processus**, ce qui réduit considérablement les coûts d'exploitation et de maintenance d'un système distribué.

🎯 Frontend (FE)

- Gestion des requêtes utilisateur
- Analyse et planification des requêtes (Query Planner)
- Gestion des métadonnées (schémas, tables, partitions)
- Gestion du cluster (coordination des nœuds)

⚙️ Backend (BE)

- Stockage des données (colonnes, partitions, répliques)
- Exécution des plans de requêtes
- Traitement parallèle massivement distribué **(MPP)**

MPP = Massive Parallel Processing

*Au lieu de traiter la requête sur un seul serveur, Doris la **découpe** et l'exécute **simultanément sur tous les nœuds Backend**. Chaque BE traite sa portion de données en parallèle, puis les résultats sont agrégés.*

*Résultat : performance **multipliée par le nombre de nœuds** !*

5 raisons d'adopter cette architecture

📈 **Scalabilité horizontale** : Ajoutez des FE/BE à chaud pour augmenter les capacités

🛡️ **Haute disponibilité** : Réplication automatique des métadonnées et données avec tolérance aux pannes

🔧 **Maintenance simplifiée** : Aucune dépendance externe (pas de Zookeeper, HDFS, etc.)

🔌 **Compatible MySQL** : Connexion via protocole MySQL standard (DBeaver, clients MySQL, etc.)

📊 **Stockage optimisé** : Format colonnes avec compression intelligente et indexation avancée

Sécurité et gestion des utilisateurs

Apache Doris intègre un **système de gestion des droits robuste**, permettant de contrôler finement l'accès aux données et de créer des environnements multi-utilisateurs sécurisés.

👥 Contrôle d'accès RBAC

- **Role-Based Access Control** : Gestion par rôles
- Contrôle d'accès au niveau des **bases de données** et **tables**
- Rôles personnalisables avec **héritage de permissions**
- Gestion granulaire : SELECT, INSERT, UPDATE, DELETE, ALTER

💻 Exemple de configuration

```
-- Créer un utilisateur
CREATE USER 'etudiant'@'%' IDENTIFIED BY 'password' ;
-- Créer un rôle “analyste”
CREATE ROLE 'analyste';
-- Donner les droits de lecture au rôle
GRANT SELECT ON db.* TO ROLE 'analyste';
-- Assigner le rôle à l’utilisateur
GRANT 'analyste' TO 'etudiant'@'%';
```

🔒 **Sécurité additionnelle :** Support SSL/TLS, authentification LDAP, et intégration avec Apache Ranger pour une gestion centralisée

Success Story :
10 000+ requêtes/jour chez Xiaomi

Comment Xiaomi utilise Apache Doris pour l'analyse temps réel à grande échelle

Depuis 2019, Xiaomi a déployé Apache Doris dans des **dizaines de départements** avec des centaines de nœuds pour répondre aux besoins croissants d'analyse de données. Doris gère désormais la **majorité des analyses en ligne** de l'entreprise.

40+
Clusters déployés à travers l'entreprise

100+
Noeuds dans le plus grand cluster

10k+
Requêtes d'analyse exécutées chaque jour

Cas d'utilisation chez Xiaomi

🔬
Plateforme A/B Test
Tests et expérimentations produit

📈
Analyse de croissance
Métriques business temps réel

💰
Analyse financière
Traitement des données financières

📊
Dashboards
Pour tous les groupes commerciaux

👤
Profils utilisateurs
Analyse comportementale des clients

📱
Publicité
Mesure des performances des campagnes

Source : doris.apache.org/blog/xiaomi_vector

Conclusion et ressources

Apache Doris en 3 points clés

✅ **Architecture MPP pour l'OLAP temps réel**
Traitement parallèle sur tous les nœuds pour des performances sub-seconde

✅ **Architecture minimaliste**
Seulement 2 types de processus (FE/BE), facile à déployer

✅ **Compatible MySQL et facilement intégrable**
Protocole MySQL standard, connexion via DBeaver, Power BI, etc.

💡 Quand choisir Apache Doris ?

- ❖ Analyses temps réel avec forte concurrence
- ❖ Dashboards et reporting interactifs
- ❖ Requêtes analytiques complexes (jointures, agrégations)

📚 Ressources pour aller plus loin

🌐 Site officiel : doris.apache.org

📖 Documentation : doris.apache.org/docs

💻 Code source : github.com/apache/doris

Plan de la démonstration

Les étapes pour mettre Apache Doris en action (~15 minutes)

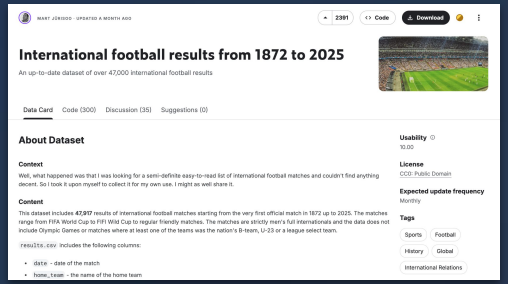
1. **Vérification de l'environnement**
   Vérification que Apache Doris (via Docker) est bien démarré et connexion via DBeaver
2. **Création d'une base de données**
   Commande SQL simple pour créer notre environnement de travail
3. **Création d'une table**
   Définition du schéma avec types de colonnes, partitionnement et modèle de données
4. **Import de données**
   Chargement d'un échantillon de données pour tester les requêtes
5. **Première requête analytique**
   Exécution d'un SELECT avec agrégation pour démontrer les capacités OLAP
6. **Visualisation des performances**
   Analyse du temps d'exécution

À vous de jouer !

⚽ **Cas pratique : Analysez 150 ans de football mondial !**

47 000+ matchs internationaux à explorer avec Apache Doris

⏱️ Durée du TD : 2 heures



😏 *Question du jour :*

Les données feront-elles ressortir la meilleure équipe et le meilleur joueur… ou vos affinités personnelles vont-elles biaiser vos analyses ?