DATA

hackA tech

Shake science. Shape innovation.

#BigData

#streaming

#systèmedistribué

KERA
Optimiser le traitement des données à grande échelle

Inria

CARACTÉRISTIQUES

Aujourd'hui les données sont devenues un enjeu stratégique majeur. Chaque jour, près de de 16.1 zettabytes de données sont générées. Pour 2025, ce volume atteindra près de 163 ZB (soit 10^{21}). Cette explosion de volume d'informations nécessite de nouveaux outils.

KerA est une plateforme permettant d'ingérer des données en continu. Celles-ci peuvent être de natures diverses (vidéos, images, textes, données structurées...), et être issues de multiples sources : capteurs, sites internet, logiciels... Elle permet d'optimiser l'échange et le stockage de données entre les sites de production et les logiciels de traitement.

KerA se concentre sur la réduction de la latence entre la lecture et la publication des données : les résultats sont proches d'être obtenus en temps réel. La plateforme est distribuée : les ressources matérielles disponibles ne se trouvent pas sur la même machine. Cette caractéristique permet d'assurer une grande montée en charge et de faciliter la réplication des données sur plusieurs serveurs afin de tolérer les pannes éventuelles.

© Inria / Photo H. Raguet

TRAITEMENT DES DONNÉES

KerA centralise les flux de données en provenance des différents sites de production d'une infrastructure, afin de limiter les problématiques d'intégration avec les logiciels de traitement. Les données ingérées par la plateforme peuvent provenir de capteurs disséminés sur le terrain (monitoring), d'évènements divers générés sur les réseaux sociaux, de simulations numériques, etc.

QUELS AVANTAGES ?

- Traitement des données en temps réel
- Tolérance aux pannes
- Facilité d'installation et de configuration
- Optimisation de la consommation d'énergie

USE CASES

- **Grille de capteurs :** champ d'éoliennes
- **Voiture autonome**
- **Réseaux sociaux :**

Exemple : le projet de startup Zettaflow consiste à adapter les avancées de KerA aux objets connectés.

FICHE IDENTITÉ

- Intégration : Apache Flink
- Langage de programmation: : C++, bindings écrits en Java & Python
- Équipe projet : KERDATA

FONCTIONNALITÉS GÉNÉRIQUES

KerA repose sur le log, une structure de données similaire à un tableau de messages au sein duquel les données sont ordonnées par date de réception. La plateforme partitionne les messages entrants selon leur catégorie (*topic*), chaque sous-partition étant un log. La capacité d'une partition peut être augmentée dynamiquement par l'ajout d'une sous-partition. Ceci permet de faciliter la configuration préalable. Chaque producteur batche (traitement de données par lots) un nombre fixe d'entrées, qui sont ajoutées à la fin d'une partition.

L'architecture de KerA est similaire à celle d'Apache Kafka : une couche de brokers traite les requêtes RPC émises par les producteurs et consommateurs. KerA apporte de nombreuses optimisations, notamment le support du *kernel-bypass* et du *zero-copy networking*. Les *brokers* gèrent la mémoire principale des serveurs en manipulant les partitions, le coordinateur gère la configuration et les métadonnées du cluster, et les *backups* gèrent la réplication de la mémoire principale.

CONNAISSANCES MINIMUM REQUISES

- **Maîtrise du C++**
- **Connaissances en systèmes distribués**
- **Connaissance de l'environnement Big Data :** problématiques, concepts, logiciels.

READ ME

Équipe projet : Kerdata https://team.inria.fr/kerdata/

Documentation publique : https://kerdata.gitlabpages.inria.fr/Kerdata-Codes/kera-website/

Thèse : https://hal.archives-ouvertes.fr/tel-01972280/

Référent : Thomas Bouvier
Kerdata est une équipe-projet commune à Inria, ENS Rennes et INSA Rennes.