Intoduction à l'architecture des systèmes distribués

Joseph Allemandou

Octobre 2016

Ce cours a pour objectif de présenter de façon pragmatique les notions nécessaires à la compréhension des architectures de systèmes distribués. Les exemples utilisés sont de complexité croissante :

— Parallélisation d'un algorithme
— Résilience et distribution de charge d'un site web
— Stockage et calcul distribué — HADOOP.

1 Jour 1 — Première approche

1.1 Calculer plus vite - Distribuer les traitements

— Exécution parallèle [1] [2] [3]
— Threads & Processus [4] [5]
— Synchronisation [6] [7]
— Appronfondissment [8] [9] [10] [11] [12] [13] [14] [15]

1.2 La résilience — L'autre bénéfice

— Architecture distribuée [16] — L'exemple d'un site web
— Bénéfices apportés par la réplication [17] [18] [19]
— Modèles de réplication [20] [21]
— Approfondissement [22] [23] [24] [25]

2 Jour 2 — HADOOP [26]

2.1 Généralités et histoire

— Passage à l'échelle [27] [28] [29]
— L'approche historique [30] [31]
— La révolution par le web [32] [33]
— À garder en mémoire [34]

2.2 Architecture du système de stockage HDFS

— Fonctionnement
— Briques logicielles nécessaires (namenode, datanode)

2.3 Architecture du système de gestion de ressources YARN

— Fonctionnement
— Briques logicielles nécessaires (resource manager, application master, worker)

2.4 Paradigme de calcul MapReduce [35]

— Fonctionnement
— Points forts et points faibles

2.5 Écosystème

Hive [36], Pig [37], HBase [38], Spark [39]

3 Jour 3 — Big Data & Évaluation

4 Vous avez dit Big Data [40] ?

En vrac ... NoSQL [41], entrpôts de données [42] [43], travail de la donnée [44] [45] [46] [47]

Références Wikipédia

[1] Parallélisme (informatique). https://fr.wikipedia.org/wiki/Parall%C3%A9lisme_(informatique).

[2] Programmation concurrente. https://fr.wikipedia.org/wiki/Programmation_concurrente.

[3] Multitâche. https://fr.wikipedia.org/wiki/Multit%C3%A2che.

[4] Thread (informatique). https://fr.wikipedia.org/wiki/Thread_(informatique).

[5] Processus (informatique). https://fr.wikipedia.org/wiki/Processus_(informatique).

[6] Synchronisation (multitâches). https://fr.wikipedia.org/wiki/Synchronisation_(multit%C3%A2ches).

[7] Interblocage. https://fr.wikipedia.org/wiki/Interblocage.

[8] Ordonnancement dans les systèmes d'exploitation. https://fr.wikipedia.org/wiki/Ordonnancement_dans_les_syst%C3%A8mes_d%27exploitation.

[9] Communication inter-processus. https://fr.wikipedia.org/wiki/Communication_inter-processus.

[10] Verrou (informatique). https://fr.wikipedia.org/wiki/Verrou_(informatique).

[11] Sémaphore (informatique). https://fr.wikipedia.org/wiki/S%C3%A9maphore_(informatique).

[12] Exclusion mutuelle. https://fr.wikipedia.org/wiki/Exclusion_mutuelle.

[13] Section critique. https://fr.wikipedia.org/wiki/Section_critique.

[14] Problème des producteurs et des consommateurs. https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_producteurs_et_des_consommateurs.

[15] Problème des lecteurs et des rédacteurs. https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_lecteurs_et_des_r%C3%A9dacteurs.

[16] Architecture distribuée. https://fr.wikipedia.org/wiki/Architecture_distribu%C3%A9e.

[17] Réplication (informatique). https://fr.wikipedia.org/wiki/R%C3%A9plication_(informatique).

[18] Point individuel de défaillance. https://fr.wikipedia.org/wiki/Point_individuel_de_d%C3%A9faillance.

[19] Répartition de charge. https://fr.wikipedia.org/wiki/R%C3%A9partition_de_charge.

[20] Maître-esclave. https://fr.wikipedia.org/wiki/Ma%C3%AEtre-esclave.

[21] Réplication multi-maîtres. https://fr.wikipedia.org/wiki/R%C3%A9plication_multi-ma%C3%AEtres.

[22] Théorème cap. https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_CAP.

[23] Consensus (informatique). https://fr.wikipedia.org/wiki/Consensus_(informatique).

[24] Paxos (informatique). https://fr.wikipedia.org/wiki/Paxos_(informatique).

[25] Split-brain. https://fr.wikipedia.org/wiki/Split-brain.

[26] Hadoop. https://fr.wikipedia.org/wiki/Hadoop.

[27] Scalability. https://fr.wikipedia.org/wiki/Scalability.

[28] Grille informatique. https://fr.wikipedia.org/wiki/Grille_informatique.

[29] Calcul distribué. https://fr.wikipedia.org/wiki/Calcul_distribu%C3%A9.

[30] Superordinateur. https://fr.wikipedia.org/wiki/Superordinateur.

[31] Openmp. https://fr.wikipedia.org/wiki/OpenMP.

[32] Grappe de serveurs. https://fr.wikipedia.org/wiki/Grappe_de_serveurs.

[33] Cluster management. https://fr.wikipedia.org/wiki/Cluster_management.

[34] Illusions de l'informatique distribuée. https://fr.wikipedia.org/wiki/Illusions_de_l%27informatique_distribu%C3%A9e.

[35] Mapreduce. https://fr.wikipedia.org/wiki/MapReduce.

[36] Apache hive. https://fr.wikipedia.org/wiki/Apache_Hive.

[37] Apache pig. https://fr.wikipedia.org/wiki/Apache_Pig.

[38] Hbase. https://fr.wikipedia.org/wiki/HBase.

[39] Apache spark. https://fr.wikipedia.org/wiki/Apache_Spark.

[40] Big data. https://fr.wikipedia.org/wiki/Big_data.

[41] Nosql. https://fr.wikipedia.org/wiki/NoSQL.

[42] Entrepôt de données. https://fr.wikipedia.org/wiki/Entrep%C3%B4t_de_donn%C3%A9es.

[43] Datamart. https://fr.wikipedia.org/wiki/Datamart.

[44] Informatique décisionnelle. https://fr.wikipedia.org/wiki/Informatique_d%C3%A9cisionnelle.

[45] Exploration de données. https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es.

[46] Intelligence artificielle. https://fr.wikipedia.org/wiki/Intelligence_artificielle.

[47] Science des données. https://fr.wikipedia.org/wiki/Science_des_donn%C3%A9es.