Introduction :

Le développement phénoménal des systèmes d'informations a conduit à une augmentation des cours en ligne d'apprentissage pour les étudiants (MOOCs). Dans ces cours, les étudiants utilisent un environnement d'apprentissage virtuel (VLE), pour simuler l'expérience d'une salle de classe réelle, un des avantages les plus intéressants est la capacité d'évaluation en temps réel, la collecte du comportement étudiant et l'interaction avec les ressources des cours, toutes ces informations engendre ce qu'on appelle le "Big Data", représentant une mine d'or pour les établissement d'enseignement, elles peuvent être utilisées pour détecter les étudiants en danger d'échec et intervenir à temps et même afin d'améliorer la qualité d'apprentissage.

Contexte :

Aujourd'hui, l'analyse de l'apprentissage est utilisée pour améliorer les performances des enseignants ainsi que la qualité du contenu pédagogique fourni aux apprenants, ainsi elle offre un aperçu global et elle aide les enseignants et les écoles ainsi que les formateur et les entreprises à prendre des décisions sur la manière de rendre l'apprentissage plus efficace pour leurs élèves et apprenants. L'analyse d'apprentissage permet un enseignement basé sur les données.

Objectifs :

Dans notre travail, nous cherchons à collecter et analyser les données recueillies dans les environnements d'apprentissage sur les apprenants, durant la période d'enseignements.
Nous explorons comment utiliser ces informations issues de l'analyse de l'apprentissage pour fournir des informations sur l'enseignement et l'apprentissage et ainsi comprendre comment rendre l'apprentissage plus efficace et quels sont les critères impactant la réussite.

Figure 1 : Principal Objectifs
Source:Open University Learning Analytics dataset paper

L'acquisition de données et l'analyse de l'apprentissage offrent la possibilité de prendre des mesures au besoin pour aider les apprenants à atteindre des objectifs d'apprentissage définis. Les données de la classe peuvent être utilisées pour faire progresser les apprenants et les motiver à atteindre leurs objectifs. Il permet ainsi aux éducateurs d'être plus réactifs aux besoins des apprenants.

Données :

Nous avons opté pour une base de données très connue intitulée Open University Learning Analytics Dataset d'une université au UK, qui offre la possibilité de se former en ligne.
La base de données contient les informations sur 22 cours, 32 593 étudiants, leurs résultats d'évaluation et leurs interactions avec le VLE (Virtual Learning Environment) représentés par des résumés quotidiens des clics des étudiants (10 655 280 entrées).

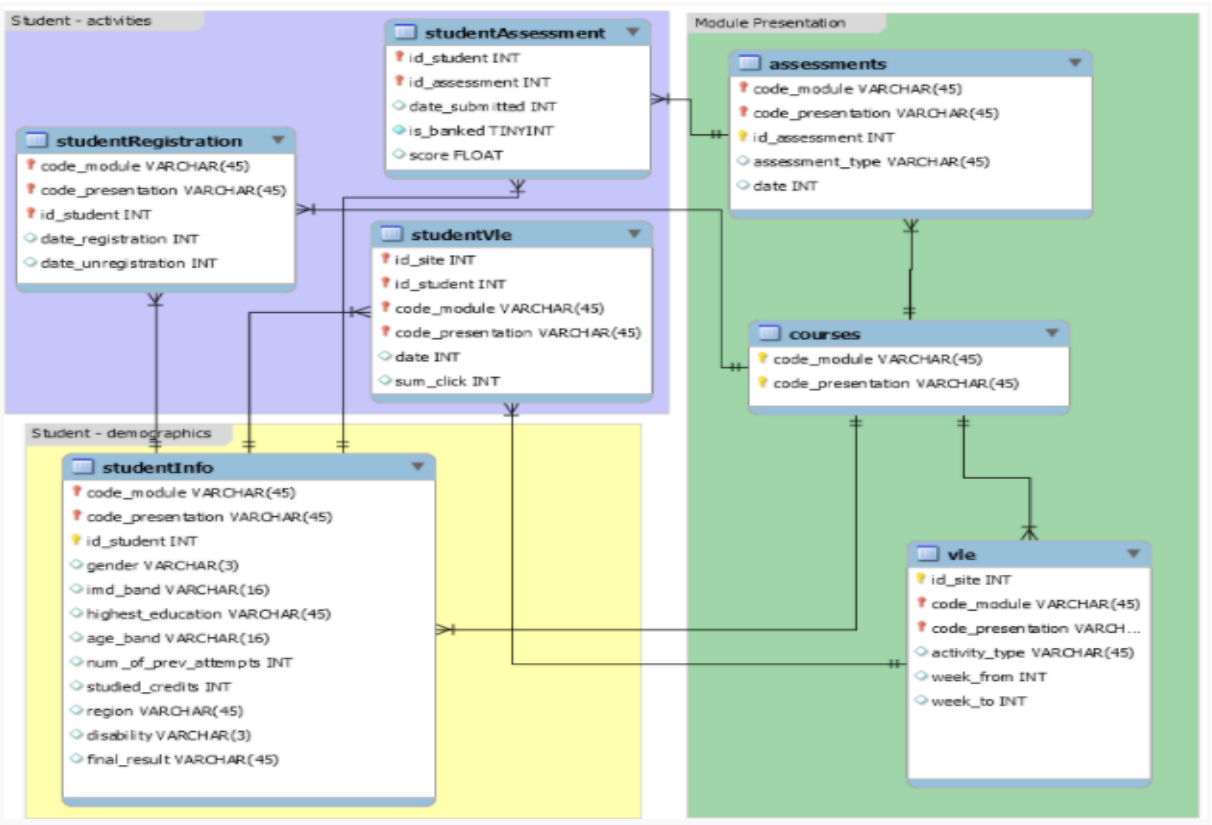


Figure 2 : Schéma de base de donnée

L'OULAD est une collection de données tabulaires sur les étudiants des années 2013 et 2014. L'ensemble de données est disponible sous la forme d'un ensemble de fichiers CSV séparés (valeurs séparées par des virgules, chaque valeur est entre guillemets et la première ligne représente les noms de colonnes).Chaque table contient des informations différentes, qui sont liées aux données d'autres tableaux à l'aide d'identificateurs uniques (colonnes). Les données contenues dans le jeu de données sont structurées comme dans la figure 2. Le jeu de données est orienté étudiant, donc l'étudiant est le point central. Les données des étudiants comprennent des informations sur leurs données démographiques et leurs inscriptions aux modules. Pour chaque triplet étudiant-module-présentation, l'ensemble de données contient les résultats des évaluations des étudiants. Les interactions des étudiants avec le VLE sont consignées sous forme de résumé de leurs activités quotidiennes.

- Démographique : représente les informations de base sur les étudiants, y compris leur âge, sexe, région, études antérieures, etc.
- Performance : reflète les résultats et les réalisations des étudiants pendant leurs études à l'Open University.
- Comportement d'apprentissage : est le journal des activités des étudiants dans le VLE.

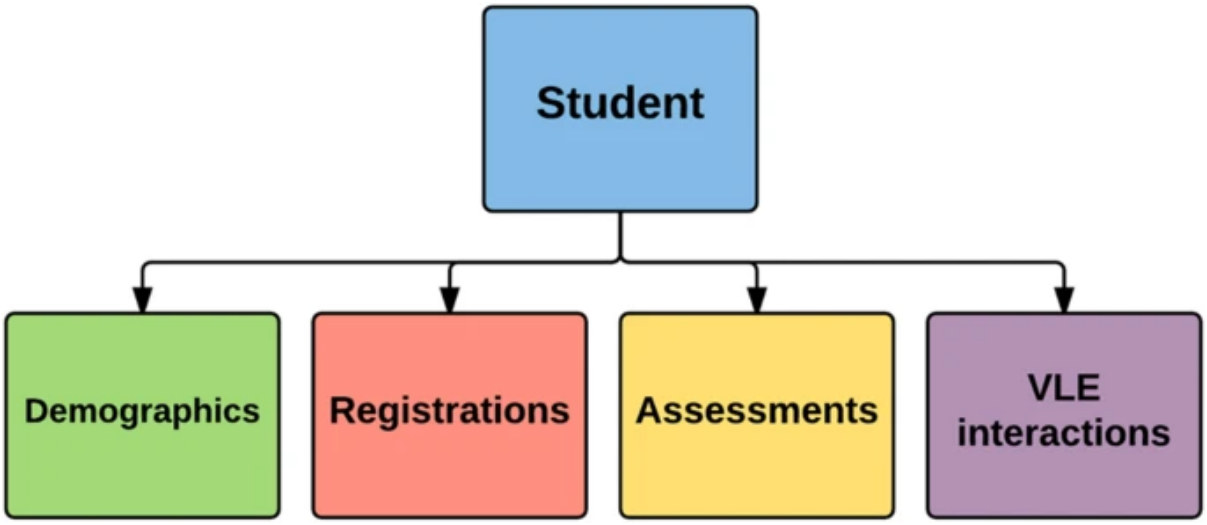


Figure 3 : Structure globale de base de donnée

Explication de certaines colonnes ambigües :

code_presentation : Les modules peuvent être présentés plusieurs fois au cours de l'année. Pour distinguer les différentes présentations d'un module, chaque présentation est nommée par l'année et le mois de son début. Par exemple, les présentations commençant en janvier se terminent par A, en février par B et ainsi de suite; de sorte que «2013 J» signifie que la présentation a commencé en octobre 2013.

date_registration : Les étudiants peuvent s'inscrire au module quelques mois avant le début de la présentation jusqu'à deux semaines après la date officielle de début du module. Chaque module comprend plusieurs évaluations. À la fin du module, il y a généralement un examen final.

assessment_type : Il existe trois types d'évaluations: l'évaluation par tuteur (TMA), l'évaluation par ordinateur (CMA) et l'examen final (examen).

Figure 4 : Structure typique du module

Architecture :

Pour la réalisation de notre projet nous avons construit un cluster avec des machines EC2 sur AWS, on a d'abord commencé par demander un compte AWS Educate qui nous a permis de bénéficier d'un crédit de 100$ afin de construire une architecture Big Data distribuée composée d'un cluster AWS EC2 composé 3 nœuds (1 NameNode 2 DataNode).

Premièrement nous avons instancié nos trois machines basé sur Red Hat Linux comme OS, chacune possédant (2 CPU, 4go de Ram, 25 go de DD).

	Family	Type	vCPUs	Memory (GiB)
☐	t2	t2.nano	1	0.5
☐	t2	t2.micro Free tier eligible	1	1
☐	t2	t2.small	1	2
☑	t2	t2.medium	2	4
☐	t2	t2.large	2	8
☐	t2	t2.xlarge	4	16
☐	t2	t2.2xlarge	8	32

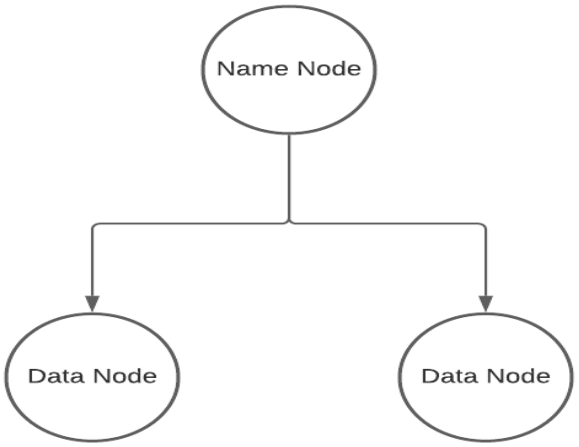
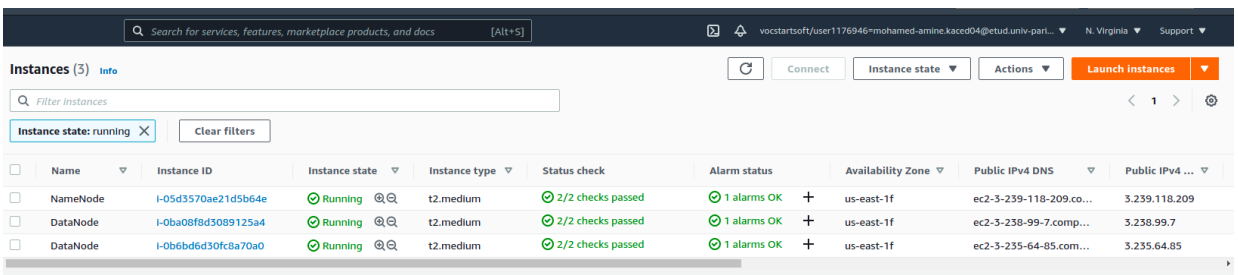


Figure 5 : Architecture du Cluster AWS



Ensuite on a récupéré la clé privée qui nous permettra d'avoir l'exclusivité de la connexion SSH.

ssh -i "AWS.pem" ec2-user@ec2-34-238-174-140.compute-1.amazonaws.com

Avec cette commande on a pu accéder à notre machine NameNode en SSH.

Cependant, il fallait encore installer et configurer l'environnement Hadoop, c'est ici que nous avons installé **cloudera-manager** qui nous a permis de configurer tous le cluster et l'écosystème, et de même installer les briques Hadoop qu'on veut.

Après, installations et configuration on a pu accéder à l'interface de cloudera manager grâce à un serveur web directement sur cette adresse :

ec2-34-238-174-140.compute-1.amazonaws.com:7180

Dans cette interface web on a la possibilité de tout gérer dans notre cluster, il propose même un dashboard qui permet le monitoring et la maintenance de notre infrastructure.

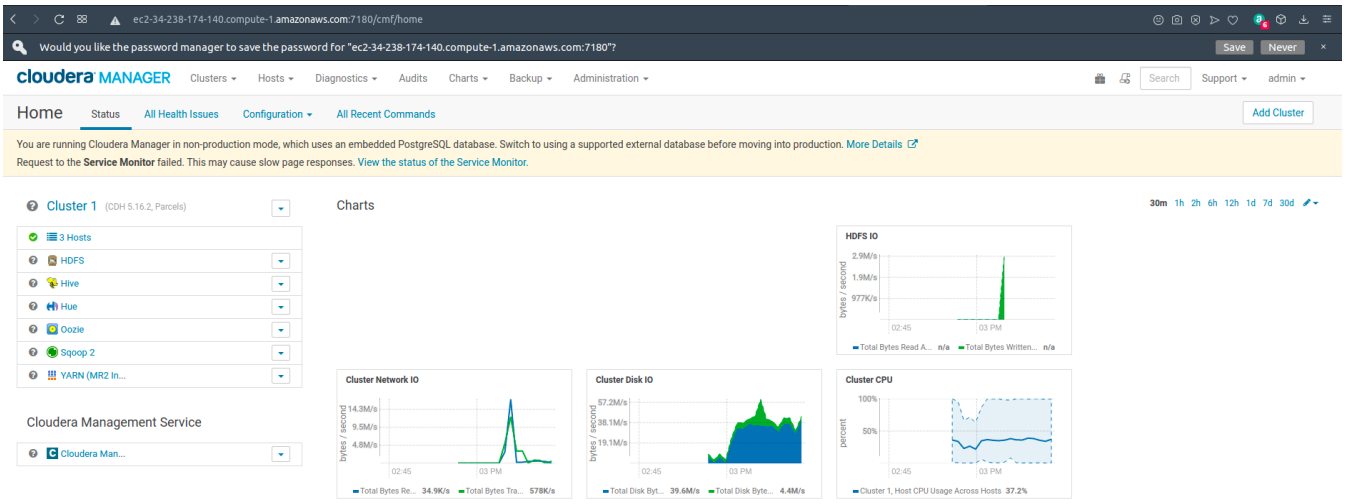


Figure 6 : Dashboard du Cluster sur Cloudera Manager

Acquisitions :

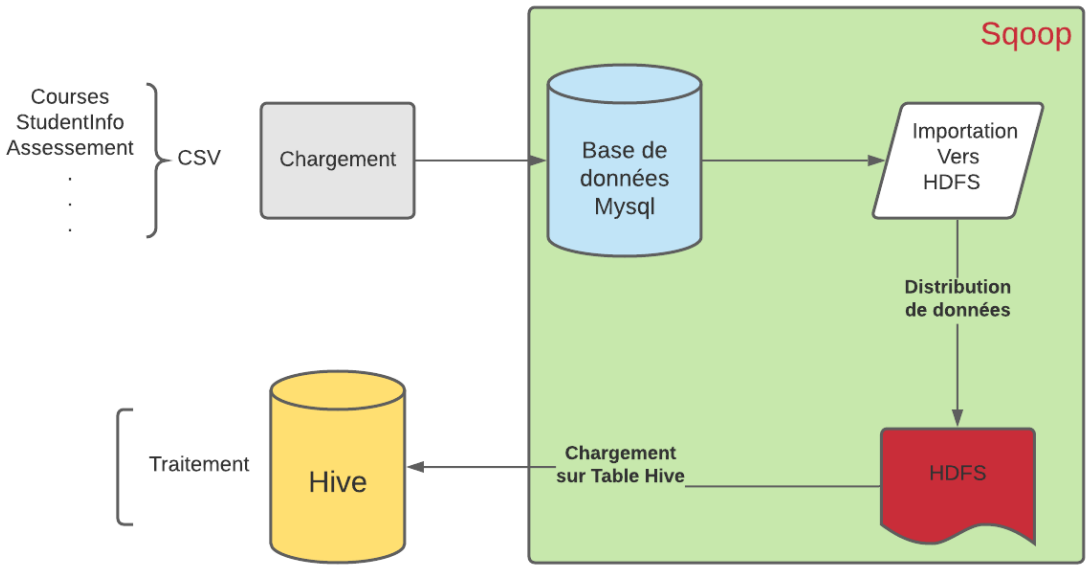


Figure 7 : Flux de données

Viens maintenant la partie transfert de données, pour se faire on a premièrement on transféré la dataset vers notre serveur avec “FileZila” en utilisant le protocole FTP, ensuite on a créé les tables et charger les données de la dataset dans la base de données relationnel Mysql car c'était son emplacement initial, en utilisant la brick SQOOP on a pu importer ces données vers HDFS sous forme distribué (en blocs), après importation ces données sont sous formes de blocs de fichiers, à leur tours ces derniers sont récupérés sur les tables externe de Hive créés déjà au préalable.

Exemple de création d’une table dans la base de donne Mysql:

create table courses (code_module varchar(20), code_presentation varchar(20), module_presentation_length int);

Ajout de la clé primaire:

alter table courses add column `id` int(10) unsigned primary KEY AUTO_INCREMENT;

Chargement des fichiers CSV:

```
LOAD DATA INFILE '/home/cloudera/Desktop/dataset/1courses.csv' INTO TABLE courses
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';
```

Importation de la table vers HDFS:

```
sqoop import --connect jdbc:mysql://localhost/students --table courses --username root -P
--target-dir students_database/courses
```

Création de la table sur base de données HIVE:

```
create external table courses ( code_module STRING,
code_presentation string, module_presentation_length int) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ",";
```

Chargement des données dans la table HIVE:

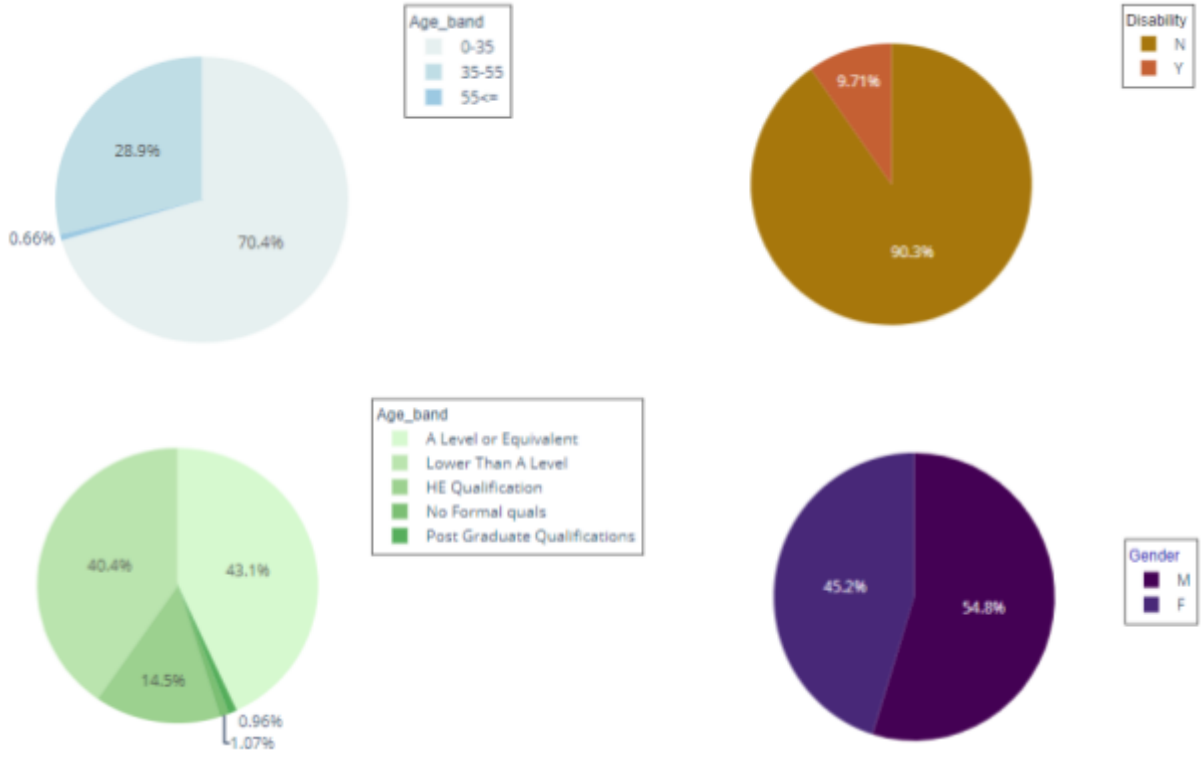
```
load data inpath "/user/cloudera/students_database/courses/part*" into table courses
```

Analyse :

Enfin on peut maintenant exécuter nos analyses et traitement des données avec Hive, où ce dernier va tirer profit de l'infrastructure distribuée et du paradigme MapReduce afin de réduire grandement le temps de traitement.

Analyse statistiques :

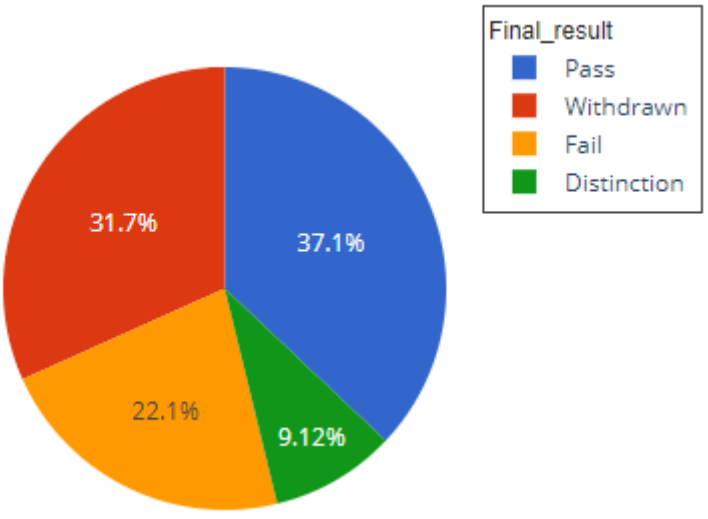
Students Profile :



Le pourcentage des résultats finaux des étudiants par sex :

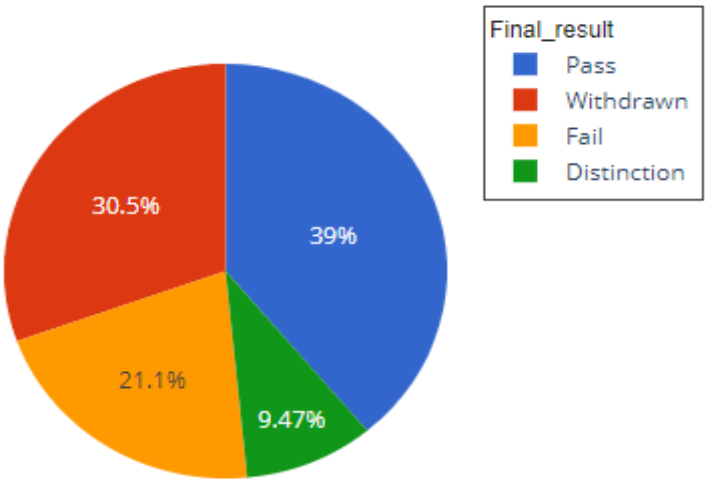
requête 1 :

```
select final_result,(count(*)/17875)*100 male_percentage
from studentinfo
where gender="M"
group by final_result
```



Requête 2 :

```
select final_result,(count(*)/14718)*100 female_percentage
from studentinfo
where gender="F"
group by final_result
```



D'après cette analyse ,on remarque qu'il n y a pas une grande différence entre le niveau des hommes et femmes en se basant sur le pourcentage des admis et les excellents donc le sex n'a pas d'impact sur le niveau intellectuel ou la réussite de l'étudiant.

Analyse prédictive :

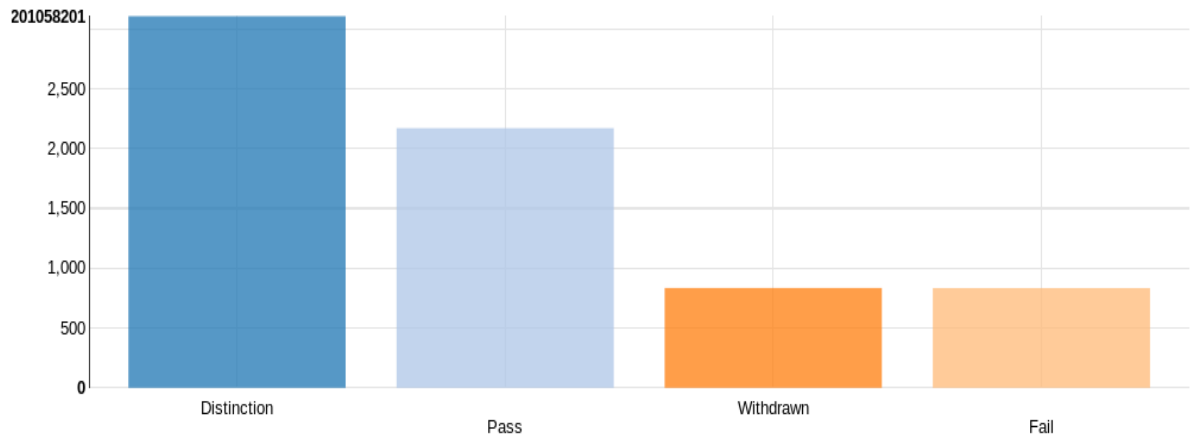
Dans cette partie on va essayer de trouver les aspects et comportements pouvant impacter le cursus d'apprentissage de l'étudiant, et utiliser ces derniers pour tirer des conclusion, qui nous permettront de pouvoir juger ou prédire les résultats d'un étudiant à partir de son comportement sur le VLE ou ses précédentes notes.

A. Le nombre total de clics a-t-il un impact sur le résultat final ?

Requête :

```
select stin.final_result, avg(click) clicks_mean
from studentinfo stin
join (select id_student,sum(sum_click) click from studentvle group by id_student) as temp
on (temp.id_student=stin.id_student)
group by stin.final_result;
```

Graphe :



Interprétation :

On remarque que les étudiants qui font plus de clics sur le matériel , obtiennent des bons résultats (excellents et admis).

Déduction :

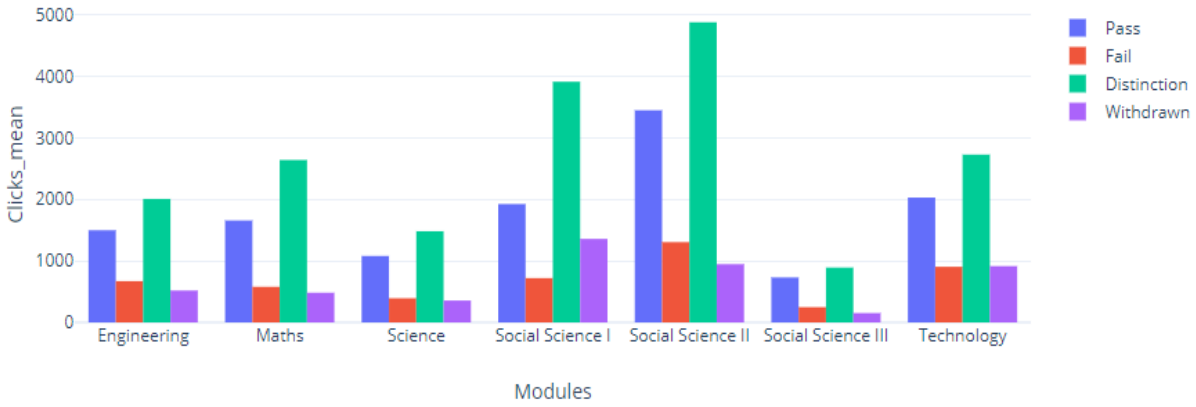
On peut déduire que l'utilisation de l'environnement virtuel peut affecter énormément les résultats finals.

B. Nombres moyens de clics pour chaque statut dans chaque module:

Requête :

```
select stin.code_module,stin.final_result,avg(click) clicks_mean
from studentinfo stin
join (select id_student,code_module,sum(sum_click) click from studentvle group by id_student, code_module) as temp
on (temp.id_student=stin.id_student)
group by stin.final_result,stin.code_module
order by code_module asc;
```

Graphe :



Interprétation :

- La validation d'un module requiert en moyenne un nombre de clics avoisinant 3000-2000.

- Pour que l'étudiant atteigne le niveau d'excellence dans un module il doit avoir en moyenne 3000 clics pour les Sciences Social, dans les restes modules il suffit d'un nombre de clics qui dépasse les 2000.

Déduction :

Les modules social science 1 et 2 nécessitent beaucoup d'interaction avec le matériel virtuel pour valider le module ou être excellent par rapport au autres modules.

C. ***Relation entre le nombre total de clics et les modules:***

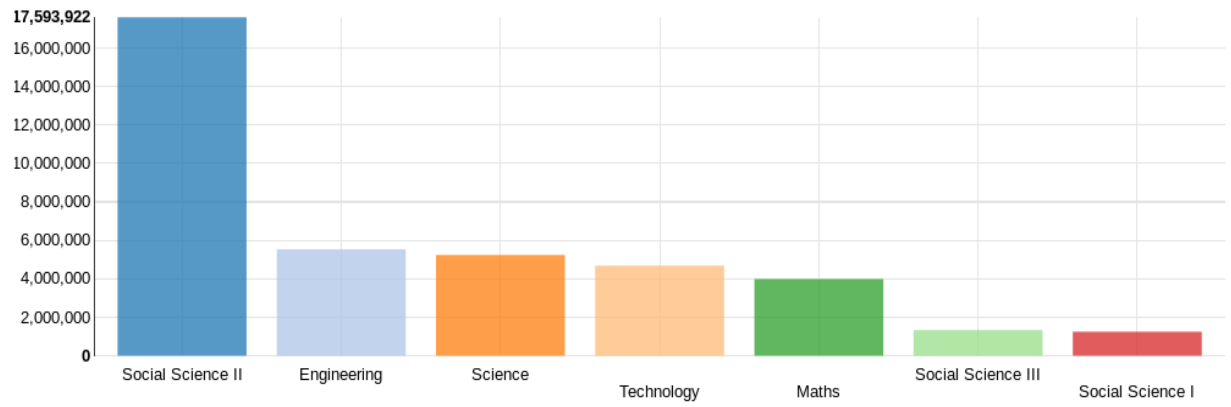
Requête :

select code_module,sum(sum_click) clicks_nbr

from studentvle

group by code_module

Graphe :



Interprétation :

On voit que les étudiants s'intéressent et interagissent beaucoup au module Social science 2 par rapport à d'autres modules.

Déduction :

Peut-être que l'enseignement de ce module est plus interactif où il a une méthode qui attire plus ses étudiants ou peut être le contenu du module est intéressant pour les étudiants.

D. ***La moyenne des clics requis pour réussir dans chaque module:***

Requête :

select stin.code_module,stin.final_result,avg(click) clicks_mean

from studentinfo stin join (select id_student,code_module,sum(sum_click) click

from studentvle group by id_student, code_module) as temp

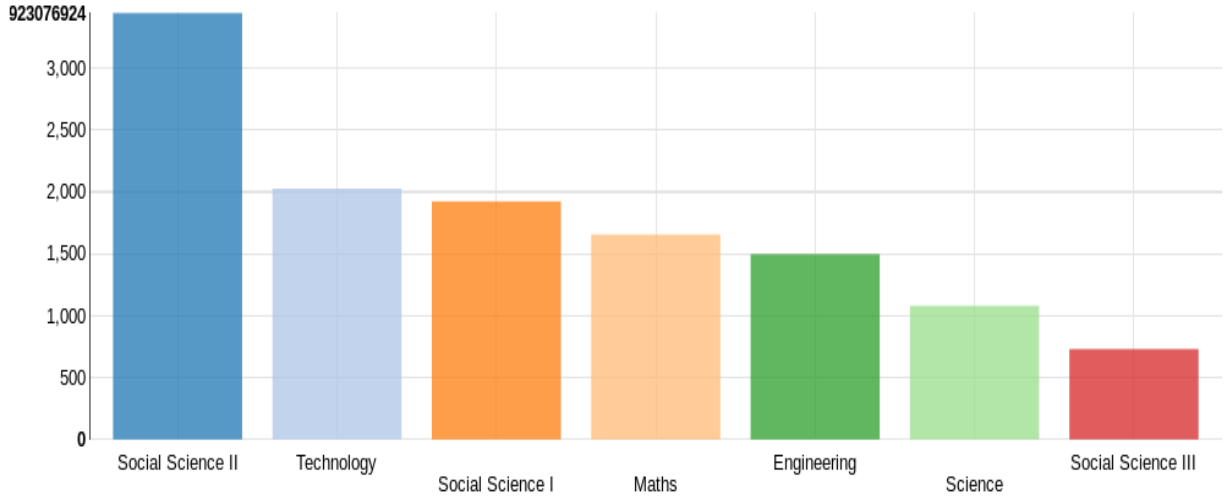
on (temp.id_student=stin.id_student)

where stin.final_result="Pass"

group by stin.final_result,stin.code_module

order by code_module asc

Graphe :



Interprétation :

La moyenne des clics pour passer dans le module social science 2 est très élevée, pour la technologie et social science 1 est peu élevé, pour réussir dans social science 3 l'etudiant n'as pas besoin d'interagir beaucoup avec l'environnement virtuel.

E. L'impact de l'handicap sur le résultat final :

Requête 1 :

```
select disability,final_result,(count(*)/29429)*100 nbr
from studentinfo
where disability ="N"
group by disability,final_result
```

Requête 2 :

```
select disability,final_result,(count(*)/3164)*100 nbr
from studentinfo
where disability ="Y"
group by disability,final_result
```

Interprétation :

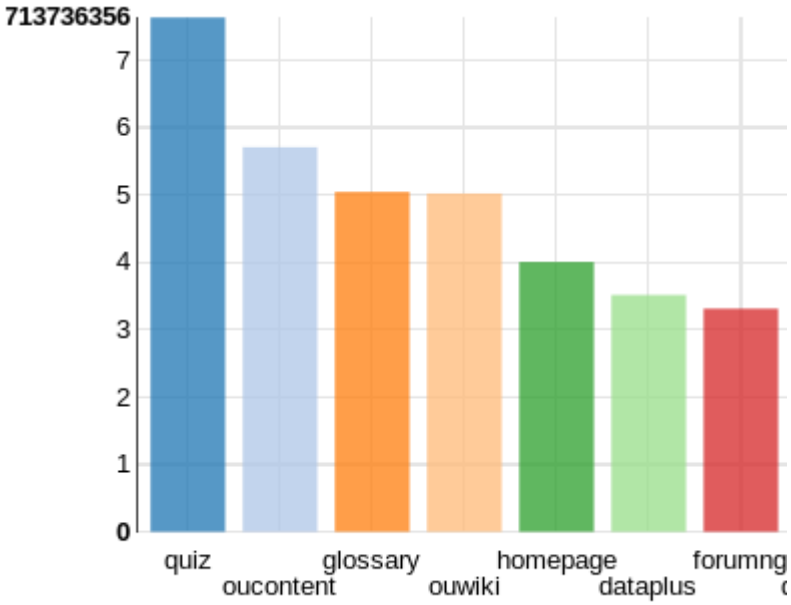
On remarque que les étudiants handicapés ont des résultats moins bons que les personnes sans handicaps , cela peut nous informer que l'handicape peut affecter le rendement et la performance de certains étudiants.

F. La moyenne des clics par activité d'apprentissage :

Requête :

```
select vle.activity_type,avg(sum_click) clicks_mean
from studentvle
where vle.activity_type in ("quiz","oucontent","glossary","ouwiki","dataplus","forumng")
join vle
on (studentvle.id_site=vle.id_site)
group by vle.activity_type
```

Graphe :



Interprétation :

On remarque que les étudiants cliquent beaucoup sur les quiz par rapport au autres types d'activités

Déduction :

les étudiants préfèrent plus des méthodes d' apprentissage divertissantes par exemple des exercices sous forme des jeux par rapport au autre types comme les cours simples.

G. Le niveau d'éducation peut-il influencer le résultat final ?

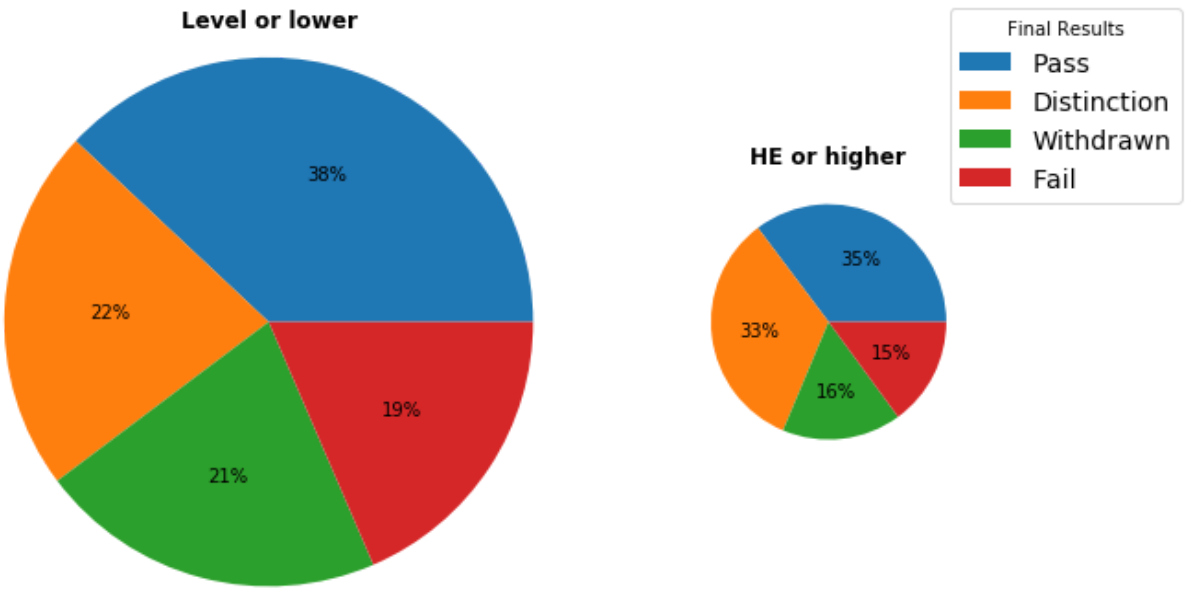
Requête 1:

select final_result ,count(*)/27203*100 from students_info where highest_education in ("A Level or Equivalent","Lower Than A Level") group by final_result

Requête 2:

select final_result ,count(*)/5390*100 from students_info where highest_education not in ("A Level or Equivalent","Lower Than A Level") group by final_result

Graphe :



Interprétation :

-Le taux des excellents passe de 22% à 33% pour un niveau d'éducation élevé.

-Le taux des admis est presque similaire.

-Le taux des abandonnés est légèrement inférieur pour le groupe du l'éducation élevée.

-Le taux des défaillants est également légèrement inférieur pour le groupe du l'éducation élevée.

Déduction :

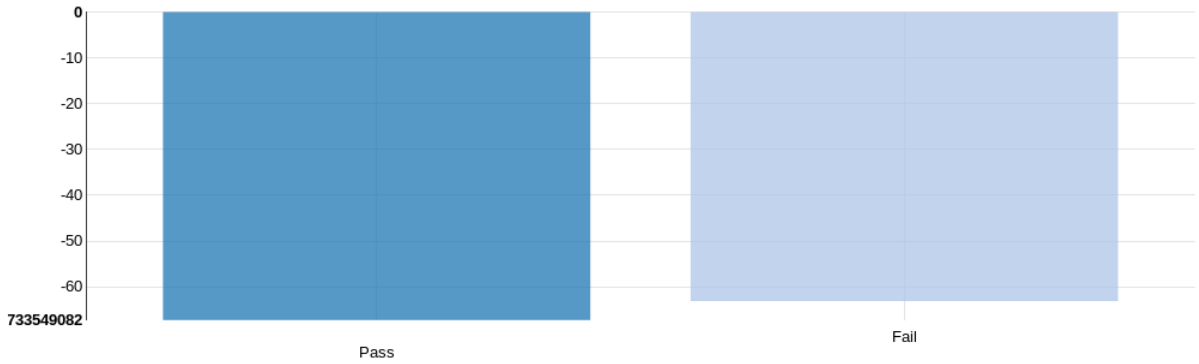
Selon ces résultats, la variable «niveau d'éducation» peut être un prédicteur du succès des élèves comme prévu par l'intuition.

H. L'impact de l'inscription tardive sur le résultat finale :

Requête :

```
select stin.final result,avg(dater) date_regis_mean
from studentinfo stin
join (select id_student,avg(date_registration) dater from studentregistration group by
id_student) as temp
on (temp.id_student=stin.id_student) where stin.final_result in ("Pass","Fail")
group by stin.final_result
```

Graphe :



Déduction :

Il est peu probable que les étudiants qui s'inscrivent tardivement aux cours réussissent que les étudiants qui s'inscrivent tôt.

Conclusion:

Bien que ce domaine est en constante évolution, il reste beaucoup de perspective prometteuse, l'une d'elle sera de mettre en place une data pipeline de logs d'interactions des étudiants sur le VLE et mettre en place un système de recommandation de cours en temps réel qui permettra d'aider les étudiants à choisir leur modules ou spécialité à partir de leur préférences, résultats et objectifs, cependant notre université devrais songer à intégrer cette technologie au sein du site moodle, celà va engendrer certainement d'énorme améliorations sur tous les niveaux.

Références

- Kuzilek, J., Hlosta, M. & Zdrahal, Z. Open University Learning Analytics dataset. Sci Data 4, 170171 (2017)

- Marie Bienkowski, Mingyu Feng, Barbara Means ,“Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief” (2012)

- Hlosta, Martin; Zdrahal, Zdenek and Zendulka, Jaroslav (2017). “Ouroboros: early identification of at-risk students without models based on legacy data. In: LAK17 - Seventh International Learning Analytics & Knowledge” Conference, 13-17 Mar 2017, Vancouver, BC, Canada, pp. 6–15.

- Murumba, Julius & Micheni, Elyjoy. (2017). Big Data Analytics in Higher Education: A Review. The International Journal of Engineering and Science. 06. 14-21. 10.9790/1813-0606021421.