Certification Data Scientist

Examen Savoir

Examen Savoir Faire

Version 2020

Veillez à bien mettre votre nom et prénom sur chaque page
Vous n'êtes pas autorisés à consulter les supports de séminaires ou vos notes pour répondre à ces QCM.
Vous devez entourer la bonne réponse.
Il n'y a qu'une seule réponse par question

Nom :

ORSYS 2020	Certification Data Scientist -- STATISTIQUES QCM SAVOIR

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

Question STA1 — **La statistique descriptive a pour objectif :**

a. de définir l'échantillon à étudier
b. de prévoir le comportement des clients
c. de décrire les données afin de mieux les analyser
d. de préparer les données dans le cadre d'un projet de datamining
e. de décrire les statistiques qui ont été réalisées sur un échantillon

Réponses			
	1		e
	2		b+c
	3		a
	4		c+d
	5		d
	6		c

Question STA2 — **La population statistique est :**

a. l'ensemble des outils mathématiques à notre disposition pour analyser les données
b. la propriété que l'on veut étudier sur l'ensemble de nos données
c. l'ensemble des objets qui sera soumis à une analyse statistique
d. les différentes modalités que peut prendre la variable étudiée
e. l'échantillon sur lequel porte notre étude

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		c+e

Question STA3 — **La langage R est :**

a. un langage formel permettant de rédiger une étude statistique
b. une bibliothèque de fonctions mathématiques
c. un langage informatique pour créer des applications dédiées à la robotique
d. un formalisme permettant de modèliser le cycle de vie des données
e. un langage informatique interprété permettant de programmer des traitements statistiques sur des jeux de données

Réponses			
	1		e
	2		a
	3		d
	4		b
	5		c
	6		b+e

Question STA4 — **La statistique descriptive est basée**

a. sur l'observation de la représentation graphique des données
b. sur la possibilité d'un échantillonnage strictement aléatoire du jeu de données
c. sur la théorie probabiliste des jeux de hasard
d. sur l'identification de la plus grande fréquence d'une occurrence
e. sur l'ensemble des études qui ont été réalisées dans le passée sur un échantillon

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		b+c

Question STA5 — **données versus information**

a. il n'y a pas de différence, les deux termes sont synonymes
b. une information est la somme de plusieurs données
c. une information est composée d'une ou plusieurs données et d'une couche sémantique
d. un jeu de données contient une information
e. l'information est le résultat d'une étude statistique descriptive d'un jeu de données

Réponses			
	1		c
	2		b
	3		d
	4		a
	5		e
	6		d+b

Question — **Dans un échantillonnage aléatoire, une donnée**

a. n'a que peu de chance d'être sélectionnée

STA6	b.	à autant de chance d'être incluse dans l'échantillon que n'importe quelle autre donnée
	c.	a une probabilité d'êre sélectionné égale à la taille de l'échantillon sur la taille de la population totale
	d.	n'a aucune chance d'être sélectionnée
	e.	est forcément sélectionnée

Réponses			
	1		d
	2		e
	3		c
	4		b+c
	5		a
	6		b

Question	**L'indexation d'un jeu de données**	
STA7	a.	n'est plus utilisé depuis l'apparition des tableurs
	b.	est automatique dans tous les tableurs
	c.	sert à conserver la temporalité des données lors de traitements statistiques
	d.	est automatique dans les logiciels statistiques ou dans le langage R
	e.	est impossible

Réponses			
	1		e
	2		d+b
	3		c+d
	4		c
	5		a
	6		b+c

Question	**La notion de distance**	
STA8	a.	n'existe pas lorsque l'on parle de données
	b.	est une notion de statistiques descriptives
	c.	est toujours la valeur absolue de la différence entre 2 données
	d.	est une notion du datamining héritée de la statistique descriptive mais peu avoir plusieurs définitions
	e.	est la mesure séparant géographiquement deux éléments de l'échantillon étudié

Réponses			
	1		a
	2		b
	3		b+d
	4		c
	5		d
	6		e

Question	**Le deuxième quartile, la moyenne et la médiane**	
STA9	a.	désignent tous les 3 la somme des données divisé par le nombre de données
	b.	sont tous les 3 des paramètres de position statistique
	c.	la moyenne et la médiane sont des paramètres de position alors que le deuxième quartile est un paramètre de dispersion
	d.	sont toujours différents
	e.	ne sont pas toujours déterminable

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		b+d

Question	**La médiane est**	
STA10	a.	l'autre nom de la moyenne
	b.	la valeur unique que deraient avoir tous les individus d'une population statistique pour que leur total soit inchangé
	c.	la valeur de la variable qui sépare l'échantillon en deux ensemble de même taille
	d.	le milieu de l'intervalle entre la valeur maximale et la valeur minimale
	e.	la valeur mitoyenne à la moyenne

Réponses			
	1		b+c
	2		a
	3		b
	4		c
	5		d
	6		e

Nom :

ORSYS 2020	Certification Data Scientist -- DATAMINING QCM SAVOIR

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

DMP1

Le datamining est ...

a. une fouille aléatoire et désorganisée au sein d'un datalake
b. l'application de fonctions mathématiques sur une population statistique
c. la description de jeux de données
d. une méthode projet disposant d'une organisation propre ainsi que d'un ensemble d'outils techniques informatiques, mathématiques et statistiques
e. un ensemble de logiciels traitant automatiquement les données

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		b+c

DMP2

On estime que le nombre de données générées dans le monde ...

a. double d'une année sur l'autre
b. sera de l'ordre de 100 zettabytes en 2020
c. correspondent à 14600g d'ADN
d. augmente de 10% par an
e. est constant

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		b+c

DMP3

La première étape d'un projet de Datamining doit être

a. la collecte de données
b. la définition de la problèmatique
c. l'achat des outils informatiques
d. le recrutement des meileurs datascientist
e. suivre une formation ORSYS

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		b+e

DMP4

L'introduction de la note méthodologique doit comporter

a. le nom de l'auteur, la date, le nom de l'entreprise
b. les objets étudiés, les dimensions, les variables et les objectifs
c. le comité de pilotage, le nom du chef de projet, les logiciels utilisés
d. les sources de données, la fréquence de rafraichissement, le coût
e. les résultats que l'on veut obtenir

Réponses			
	1		e
	2		d
	3		c
	4		b
	5		a
	6		a+e

DMP5

Les familles des différentes méthodes de Data Mining sont

a. les méthodes d'association
b. les méthodes d'analyse de séquence
c. les méthodes de clusters
d. les méthodes de classification
e. les méthodes prédictives

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		a+b+c+d+e

La spécificité dans le management d'un projet de Data Mining réside

a. dans la durée pouvant être longue avant d'obtenir un résultat

DMP6		b.	dans une progression par itérations en Z avec des changement d'outils, de paramètres et/ou de méthodologies entre chaque itération
		c.	dans sa méthode AGILE
		d.	dans l'absence de méthode spécifique
		e.	dans une méthode empirique
Réponses	**1**		e
	2		d
	3		c
	4		b
	5		a
	6		a+e

DMP7	**Les phases d'un projet de Datamining sont :**		
		a.	Préparation des données (statistique descriptive, management de la Qualité, ...)
		b.	Modélisation
		c.	Evaluation du modèle
		d.	Test
		e.	Interprétation
		f.	confrontation avec l'expert métier
		g.	déploiement
Réponses	**1**		a
	2		b
	3		c
	4		d
	5		e
	6		f
	7		g
	8		a+b+c+d
	9		a+b+c+d+e+f+g

DMP8	**Les méthodes de classification**		
		a.	divisent les objets en paquets de même taille
		b.	regroupent les objets dans des familles dont les membres ont des caractèristiques identiques
		c.	répartissent les objets par échantillonage aléatoire
		d.	classent les objets par valeurs croissantes des variables
		e.	comptabilisent les modalités de chaque variable
Réponses	**1**		e
	2		d
	3		c
	4		b
	5		a
	6		a+c

DMP9	**La notion de distance en Datamining est essentielle...**		
		a.	elle est strictement définit par une fonction mathématique
		b.	elle dérive de la notion de distance de la statistique descriptive
		c.	le data scientist peut définir sa propre notion de distance
		d.	ne détermine que la dispersion des données
		e.	n'est pas toujours définissable
Réponses	**1**		a
	2		b
	3		c
	4		d
	5		e
	6		b+c

DMP10	**Les méthodes prédictives ...**		
		a.	sont des méthodes supervisées
		b.	estiment la valeur d'une variable continue grâce à un modèle
		c.	sont basées sur un modèle construit à partir d'un catalogue pré-existant
		d.	sont basées uniquement sur des équations modélisant des mécanismes naturels vérifiés
		e.	sont basées sur les évenements ayant la plus grande fréquence
Réponses	**1**		a
	2		b
	3		c
	4		d
	5		e
	6		a+b+c

Nom :

ORSYS 2020	Certification Data Scientist -- Qualité des données QCM SAVOIR

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

MQD1

Pour une entreprise les données sont

a. le résultat d'une activité
b. éphémères et pertinentes que durant la phase de production
c. un outil décisionnel
d. un élément du capital de l'entreprise
e. une chaine de production parallèle au cœur de métier

Réponses	**1**		a
	2		b
	3		c
	4		d+a
	5		e
	6		e+d+c

MQD2

Les phases de qualité de la données sont

a. la qualité d'intégration des sources
b. la qualité des processus de saisie
c. la qualité des processus de supervision
d. la qualité des processus de transformation
e. la qualité de l'interprétation des résultats

Réponses	**1**		e
	2		d+a
	3		c+e
	4		b+c+a
	5		a+e
	6		a+b+c+d

MQD3

Les indicateurs de valeurs de l'entreprise pour le projet Qualité des données …

a. sont les Key Activity Products permettant de suivre la qualité des données de production
b. sont les Key Value Indicator basés sur les valeurs économiques
c. sont les Key Performance Indicator centrés sur les données
d. sont les Key Return Invest mesurant le retour sur investissement
e. sont les Key Data Indicator déterminant les données clées

Réponses	**1**		a
	2		b
	3		c
	4		d
	5		e
	6		b+c

MQD4

Les axes majeurs de la RGPD sont

a. l'obligation de supprimer le nom
b. la possibilité de demander à ce que ses données soient effacées
c. la possibilité d'actualiser ses données
d. la non transmission des données à un tiers
e. l'absence de sauvegarde des données

Réponses	**1**		a
	2		b+c+d
	3		c+d
	4		d+a
	5		e
	6		b+c

MQD5

Motivations d'un chantier Qualité des données

a. motivations économiques
b. obligations réglementaires
c. normes professionnelles
d. budget disponible
e. disponibilités des personnels

Réponses	**1**		a+b+c
	2		a+b
	3		b+c
	4		c
	5		d
	6		e

MQD6 — **Le BIG DATA**

a. la qualité des données n'est plus une obligation
b. ajoute de l'hérérogénité dégradant la qualité
c. augmentation non maitrisé des volumes
d. faibles coûts de la données
e. offre de nombreux services pour la qualité

Réponses	**1**		e
	2		b
	3		c
	4		b+c
	5		a
	6		b

MQD7 — **Le chantier qualité des données s'est**

a. un cycle d'amélioration continue
b. contraindre, controler, sanctionner, décider
c. planifier, déployer, controler, agir
d. cadrer, soustraiter, contractualiser, évaluer
e. distribuer, controler, déléguer, récompenser

Réponses	**1**		a
	2		b+c
	3		c
	4		d
	5		e
	6		a+c

MQD8 — **L'audit de la qualité des données …**

a. outil de profilage
b. extraction des données du SI
c. utilisation de la statistiques descriptives
d. resaisir
e. doublonner les bases de données

Réponses	**1**		e
	2		d
	3		c
	4		b
	5		a
	6		c+a

MQD9 — **Le seuil de la qualité des données est**

a. imposé par l'exactitude avec la donnée d'origine
b. relatif au besoin du métier
c. conforme aux dimensions technique
d. définit par le besoin
e. plusieurs niveaux de qualité sur la même donnée

Réponses	**1**		a
	2		b+c+d+e
	3		c+d
	4		d+e
	5		a+e
	6		d+a

MQD10 — **Définir les dimensions mesurables de la qualité**

a. proximité avec le monde réel
b. validité en format et en définition
c. intégrité avec les autres données du SI
d. complétude des données
e. utiliser les dimensions disponibles dans la littérature

Réponses	**1**		a+b+c+d
	2		b+c
	3		c+d
	4		a+e
	5		d+e
	6		e

Nom :

ORSYS 2020	Certification Data Scientist -- BIG DATA QCM SAVOIR

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

Question	Rép.	Question
BID_1		**Dans les propositions suivantes, quelles sont celles pour lesquelles le big data apporte une solution efficace :**
	1	a + c + e
	2	a + b + d + e
	3	e
	4	Ces 5 propositions sont adaptées au big data
	5	b + d mais au delà de 10 Tbytes

BID_2		HDFS est :
	1	Une base de données orienté Big Data
	2	Un composant Big Data de gestion de cluster
	3	Une distribution Big Data complète
	4	Un système de fichier dédié au Big Data
	5	Une extension du système de fichier Linux pour gérer de grosses volumétries

BID_3		Le NoSQL est un format de base de données qui se caractérise comme :
	1	a
	2	b
	3	c + d
	4	b + c + d
	5	Uniquement dédié au Big Data

BID_4		Sur une architecture Hadoop, les fichiers sont stockés grâce à :
	1	a + e
	2	b + c
	3	d + e
	4	c + d + e
	5	a + e

BID_5		Pour Hadoop, l'usage de disques en configuration RAID peut être envisagé :
	1	Aucun
	2	a
	3	a + b
	4	c
	5	a + b + c

BID_6		Le processus MapReduce est :
	1	Un mécanisme d'accélération des calculs sur les nœuds contenant les données
	2	Le mécanisme de distribution des calculs permettant le traitement en parallèle
	3	Un mécanisme en 2 étapes : 1) Chargement des données sur le MasterJob à partir de chaque dataNode - 2) Renvoi du résultat vers chaque dataNode
	4	Un mécanisme en 3 étapes : 1) Chargement des scripts par le MasterJob sur chaque dataNode - 2) Renvoi du résultat par chaque dataNode - 3) Assemblage et tri des résultats
	5	Une technologie de compression des données pour accélérer les traitements

BID_7		L'apport de HDFS 2 par rapport à la version 1 :
	1	a + b
	2	a + c + e
	3	c

	4		a+b+e
	5		b+d

BID_8		Dans l'environnement Big Data :	
	1		a+b
	2		a+b+e
	3		a
	4		a+b+c
	5		c+d+e

BID_9		Spark est un composant Big Data qui :	
	1		a
	2		b
	3		a+e
	4		b+c+d
	5		e

BID_10		Le composant Flume :	
	1		a
	2		b
	3		b+c
	4		a+c
	5		d

La société **SAFESECURE** est une PME de 250 personnes qui fabrique et commercialise des serrures, coffres forts et systèmes d'alarme pour les entreprises et les particuliers. Le siège de SAFESECURE est situé à Orléans, elle possède une cinquantaine de boutique en France ainsi qu'un réseau d'installateurs et de distribution.

La société Safesecure vient d'acquérir la société **SURV24** spécialisée dans la surveillance video principalement développée dans la région Aquitaine-Occitanie. SURV24 possède 5 centres opératonnels à Agen, Angoulême, Bordeaux, Bayonne et Niort. Le siège social est à Bordeaux. La nouvelle structure s'appelle désormais **SafeSecure24**.

L'objectif de la direction de la nouvelle SafeSecure24, est de développer une activité de surveillance sur la base du savoir faire de SURV24. Elle souhaite disposer à l'échelle nationale d'une offre allant de l'installation des solutions à leur exploitation 24x7. Le DG, Laurent Barre, est persuadé que la mise en place d'une stratégie données permettra d'assister la stratégie d'entreprise pour son développement et l'optimisation de ses coûts. Il a développé cette stratégie avec son directeur marketing Simon Ventoux. Tous deux sont sponsors actifs de mise en place de méthodes et d'outils pour développer l'analyse des données.

La société Safesecure est dotée d'une direction des systèmes d'information (DSI) dirigée par Annie Cole. Celle-ci devient DSI du nouvel ensemble. Elle a été chargée par la direction de mettre en place l'opérationnel pour développer la stratégie données de SafeSecure24.

Annie Cole a créé un poste de Data Manager qui lui est directement rattaché. Daniel Manage, qui prend ce poste, était dans l'entreprise en tant que responsable de développement produit.

Deux personnes vont rejoindre l'équipe Data Management :

- Paul Dacalté, ancien responsable qualité développement au sein de la DSI
- Jean Desdonnet, analyste décisionnel qui vient de suivre un cycle de formation Data Science. Il prend la fonction de data scientist au sein du pôle DM, rattaché à Daniel Manage.

Le comité de gouvernance des données a mandaté Annie Cole de procéder à un état des lieux de la qualité des données, la capacité de croiser les données Safesecure et SURV24

L'équipe DM se met donc au travail avec les objectifs suivants :

- Recenser les tableaux d'exploitation statistiques de données existant chez SURV24 et SafeSecure.
- Premier état des lieux général sous 2 mois des données clés et de leur état
- Etude d'un modèle cible des données consilidées et feuille de route pour sortir des exploitations statistiques SafeSecure24

Les pôles de données hérités des deux entreprises :

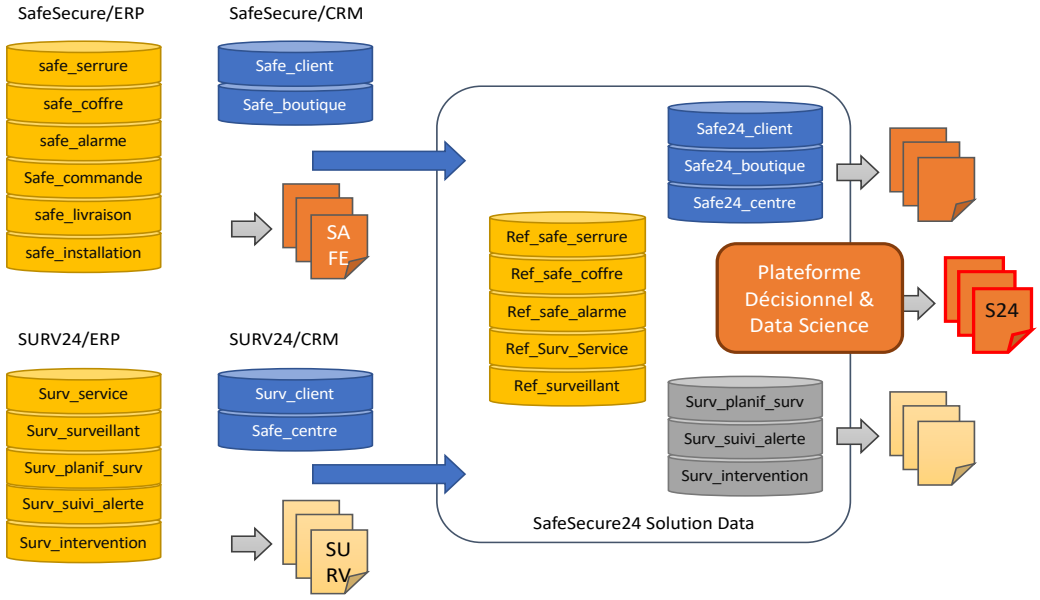
Côté SafeSecure

1. Référentiel des produits « safe_serrure » (Source SafeSecure/ERP)
2. Référentiel des produits « safe_Coffre_fort » (Source SafeSecure/ERP)
3. Référentiel des produits « safe_alarme » (Source SafeSecure/ERP)
4. Référentiel des clients « safe_client » (Source SafeSecure/CRM)
5. Référentiel des boutiques « safe_boutique » (Source SafeSecure/CRM)
6. Suivi des commandes « safe_commande » (Source SafeSecure/ERP)
7. Suivi des livraisons « safe_livraison » (Source SafeSecure/ERP)
8. Suivi des installations « safe_installation » (Source SafeSecure/ERP)

Côté SURV24

1. Référentiel des services « surv_service » (Source SURV24/ERP)
2. Référentiel des clients « surv_client » (Source SURV24/CRM)
3. Référentiel des centre de surveillance « surv_centre » (Source SURV24/CRM)
4. Référentiel des personnels d'encadrement et de surveillance « surv_surveillant » (Source SURV24/ERP)
5. Planning des « surv_planif_surv » (Source SURV24/ERP)
6. Base de suivi service et alertes « surv_suivi_alerte » (Source SURV24/ERP)
7. Base de suivi des interventions « surv_suivi_intervention » (Source SURV24/ERP)

La plateforme SafeSecure24 Solution Data : **S24SD**



Etat des lieux de l'équipe Data Management

Le résultat de l'analyse des données effectuée dans le délai de 2 mois imparti (bravo à l'équipe) n'est pas brillant :

- Côté SafeSecure et côté SURV24, les référentiels sont partiellement existants, mal gérés et ne sont pas alignés
- L'analyse des bases de données a montré de nombreux problèmes de complétude, de conformité et d'exactitude des données
- Il existe un nombre important de traitement locaux effectués avec Excel
- L'intendance de données dans les pôles métier n'existe quasiment pas et aucun indicateur n'a été mis en place

En revanche, il existe une bonne adhésion à l'analyse des données et une forte attente sur les analyses de marché et le suivi de la performance de service de la part des équipes commerciales et de production.

Proposition de consolidation des données de l'équipe Data Management

Pour produire des données consolidées, les membres de l'équipe Data Management étaient d'accord pour aller vers un rapprochement des données dans un espace technique commun : la solution de SafeSecure Solution Integration Data.

Ils n'étaient en revanche pas tout à fait en phase sur les étapes de consolidation de la SSID et au final de l'approche tout Hadoop ou non :

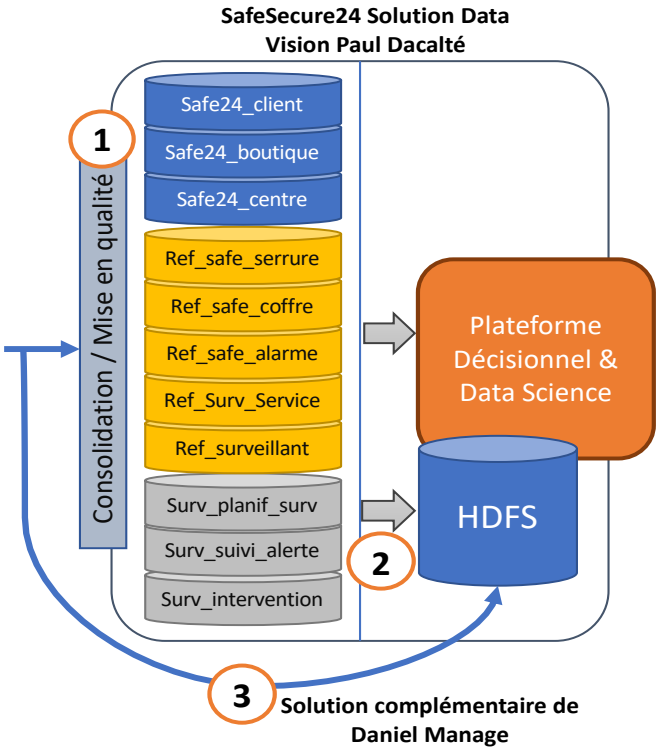
- Jean Desdonnet était en faveur de centraliser l'effort de consolidation autour d'une plateforme

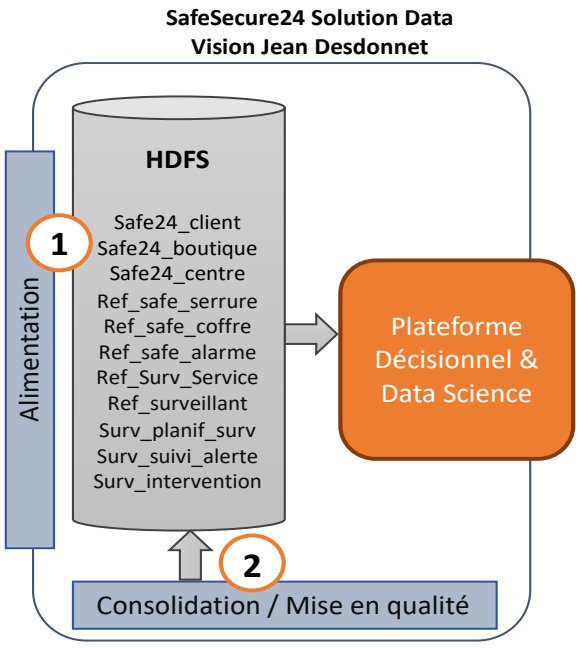
Big Data sans étape intermédiaire de consolidation des référentiels.

- Etape 1 : alimenter les données dans Hadoop
- Etape 2 : Effectuer le travail de redressement dans l'environnement Big Data

- Paul Dacalté militait lui pour la constitution d'un pôle de données redressées basées sur des référentiels partagés en amont de la Plateforme Décisionnelle et Data Science
  - Etape 1 : Effectuer redressement des données et consolidation des données pour constituer de nouveaux référentiels
  - Etape 2 : Verser dans Hadoop les données déjà consolidées

Daniel Manage a tranché pour l'approche de Paul Décalté tout en laissant une option (3) de verser directement dans Hadoop des données auxquelles on n'appliquerait aucun traitement préalable





Nom :

ORSYS 2020	Certification Data Scientist -- STATISTIQUES QCM SAVOIR-FAIRE

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

Question **La variance est**

SF-STA_1

a. une mesure de la dispersion des données
b. la moyenne des distances au carré entre chaque valeur à la moyenne
c. l'écart entre la valeur maximale et la valeur minimale des données
d. la moyenne des données divisée par la taille de la population statistique
e. le changement de typage de la variable

Réponses			
	1		a
	2		a+b
	3		b
	4		c
	5		d
	6		e

Question **L'écart type ...**

SF-STA_2

a. est la racine carrée de la variance pour avoir la même dimension que les données
b. est une mesure de la dispersion des données
c. mesure la distance entre 2 individus de la population statistique
d. l'intervalle optimale entre les données afin d'avoir le meilleur intervalle de confiance
e. la distance euclidienne entre la médiane et la moyenne

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		a+b

Question **En statistique descriptive une courbe de distribution ...**

SF-STA_3

a. définit une probabilité d'obtenir une valeur donnée
b. repose sur la moyenne et l'écart type du jeu de données
c. représente la variation des données dans le temps
d. permet de préparer les données dans le cadre d'un projet de datamining
e. de décrire les statistiques qui ont été réalisées sur un échantillon

Réponses			
	1		a
	2		b
	3		b+a
	4		c
	5		d
	6		e

Question **La covariance ...**

SF-STA_4

a. est la moyenne du produit des distances de chacune des 2 variables à leur moyenne respective
b. est la variance divisé par l'écart type
c. décrit les données afin de mieux les analyser
d. indique si 2 données ont des variations similaires dans le temps
e. permet de déterminer si l'une des données dépend linéairement d'une autre données

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e+a
	6		c

Question **Le coefficient de corrélation linéaire est**

SF-STA_5

a. la covariance divisé par le produit des écarts types
b. indique la pente de la droite reliant 2 variables entre elles
c. permet de passer d'une variable à l'autre en applicant le coefficient de corrélation
d. définit une droite passant au plus prêt des valeurs
e. permet de tracer une courbe extrapolant les données manquantes

Réponses			
	1		a+b
	2		b+d
	3		c
	4		d
	5		e
	6		d+e

Nom :

ORSYS 2020	Certification Data Scientist -- DATAMINING QCM SAVOIR-FAIRE	ORSYS formation

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

SF-DMP1

Les méthodes supervisées…

a. ne sont pas utilisées pour prendre des décisions
b. affectent un objet dans une classe en utilisant un modèle
c. utilisent un modèle construit à partir d'un catalogue d'objet pré-existant
d. effectuent un classement basé une action réalisé par un opérateur
e. classent un objet dans plusieurs classes différentes

Réponses			
	1		a+d
	2		e+d
	3		d+a
	4		b+c
	5		b
	6		e

SF-DMP2

Les méthodes non-supervisées

a. n'utilisent pas de modèle
b. créent des classes en utilisant une notion de distance
c. sont moins fiables que les méthodes supervisées
d. ne sont pas évolutives
e. enrichissent le modèle au fur et à mesure des itérations

Réponses			
	1		e
	2		d
	3		c
	4		b
	5		a
	6		b+e

SF-DMP3

Le textmining permet de classer un document au sein d'une collection …

a. il s'agit d'une méthode supervisée
b. il s'agit d'une méthode non-supervisée
c. en se basant sur un modèle construit à partir d'un catalogue préexistant de signature de texte
d. en fonction de la fréquence de mots clés
e. en fonction de l'auteur

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		d+e

SF-DMP4

La table de contingence est …

a. une regroupement de vecteurs basés sur la fréquence des mots du texte
b. un tableau construit par croisement de classes prédéfinies et de mots clés
c. une grille de lecture des textes permettant de classer un texte en fonction de sa proximité avec la grille de référence
d. une liste de signature de texte type
e. une comparaison de texte d'auteurs de référence

Réponses			
	1		c
	2		b
	3		d
	4		a
	5		e
	6		d+e

SF-DMP5

L'arbre de décision est

a. une représentation de causes à effets
b. une méthode de classification supervisée
c. un tirage aléatoire d'objets de notre jeu de données
d. une méthode non supervisée de segmentation
e. une règle d'association

Réponses			
	1		a
	2		b
	3		c
	4		d+e
	5		e
	6		d+e

Nom :

ORSYS 2020	Certification Data Scientist -- Qualité des données QCM SAVOIR-FAIRE

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

SF-MQD_1

La vision globale d'architecture est recommandée à une approche de la revue qualité. Paul Dacalté et Daniel Manage vont sollicité Annie Cole pour avoir ces entrants

a. La DSI demande à Daniel Manage de contribuer à cette formalisation de l'architecture
b. L'équipe Data Management peut faire une partie du travail
c. C'est au Data scientist, Paul Desdonnet, de faire cette revue
d. L'approche Big Data évite de faire ce travail
e. Daniel Manage préfère passer cette étape pour fournir plus rapidement une synthèse qualité

Réponses			
	1		a+b
	2		a
	3		c
	4		a+e
	5		d+e
	6		e

SF-MQD_2

Le cycle de vie de la donnée

a. la donnée a plusieurs cycles de vie
b. le point de départ du cycle de vie est l'évènement à l'origine de l'acquisition de la donnée
c. le cycle de vie de la donné est asservie par les indicateurs de qualité
d. le positionnement des indicateurs n'est pas important
e. il faut placer une mesure des indicateurs à chaque action

Réponses			
	1		a
	2		b+c
	3		b
	4		c
	5		d
	6		e

SF-MQD_3

Les indicateurs de Qualité

a. plus il y en a, mieux c'est
b. ils doivent être pertinent
c. ils doivent être facilement mesurables
d. un positionnement en début et en fin de cycle de vie suffisent
e. la convention de calcul de l'indicateur n'est pas importante

Réponses			
	1		a
	2		b+c
	3		c
	4		d
	5		e
	6		b

SF-MQD_4

Paul Dacalté organise un atelier de définition des indicateurs qualité avec le métier. Il prévoit de définir pour chaque indicateur :

a. complétude, disponibilité, sécurité
b. quantité, temps, coût
c. objectifs, données concernées, définition
d. conformité, objectifs, disponibilité
e. méthode de calcul

Réponses			
	1		a
	2		b
	3		c
	4		d
	5		e
	6		c+e

Annie Cole demande à Daniel Manage de définir la Maturité de l'entreprise :

SF-MQD_5		a.	Avec l'identification du défaut de qualité, l'entreprise a la maturité nécessaire pour entamer un projet démarche qualité
		b.	La maturité de l'entreprise est atteinte dès la définition des dimensions de la qualité
		c.	Les règles de supervision des processus n'interviennent pas dans le niveau de maturité
		d.	La définiton et la supervision de dimensions de la qualité constituent une première étape dans la maturité de l'entreprise
		e.	Une gouvernance de la qualité révèle un fort niveau de maturité de l'entreprise
Réponses	**1**		a
	2		b
	3		c
	4		d+e
	5		e
	6		d

SF-MQD_6	**La DSI demande à Daniel Manage quels vont être ses besoins. Les outils pour la qualité des données :**		
		a.	Il n'existe pas d'outils pour mesurer et suivre la qualité des données
		b.	les ETL de la suite décisionnels peuvent être utilisés pour mesurer et suivre le projet Qualité des données
		c.	La statistiques descriptives offrent de nombreux outils permettant de définir des indicateurs de la qualité des données
		d.	Des outils open sources en particulier pour le datamining permettent de mesurer la qualité des données dans le temps.
		e.	Un simple tableur ne suffit pas pour mesurer la qualité des données
Réponses	**1**		a
	2		b+c+d
	3		c+d
	4		d+e
	5		e
	6		a+e

Nom :

ORSYS 2020	Certification Data Scientist -- QCM SAVOIR-FAIRE BIG DATA Basé sur étude de cas SafeSecure24

Entourer la case réponse (une seule) contenant la bonne proposition ou la combinaison la plus complète de bonnes propositions

Question	Rép.		Question
SF-BID 1			Jean Desdonnet envisage la plateforme Big Data comme une solution générique pour collecter toutes les données et appliquer l'ensemble des traitements : a - Le Big Data va en effet lui permettre d'intégrer des sources de formats divers b - Le Big Data permettra d'intégrer des sources Open Data c - Le Big Data est pauvre en librairies de calcul statistique d - Les données générées par le Big Data vont être difficilement intégrables aux rapports de pilotage e - Le big Data est surtout utile pour analyser les log applicatifs, il va être déçu dans l'usage qu'il peut en tirer
	1		a+e
	2		a+b+c
	3		a+d+e
	4		c+e
	5		a+b
SF-BID 2			Il n'existe pas de compétences Big Data au sein des équipes SafeSecure24. Paul Desdonnet argumente qu'en s'appuyant sur une distribution Cloudera, il est possible d'envisager en un mois le lancement d'une première plateforme de démonstration. Tout le monde n'est pas d'accord avec cette approche. Quelles propositions pourriez vous soutenir / Recommander : a - Vous êtes d'accord avec Paul Desdonnet b - Annie Cole, la DSI, pense qu'il faut trois mois pour mettre à disposition une plateforme Cloudera et six mois pour la production c - Annie Cole pense qu'il faut ré-écrire totalement les traitements statistiques existants pour les porter sur la nouvelle plateforme d- Paul Desdonnet affirme qu'il peut lancer des premiers développements sur une plateforme Cloudera de test avec des jeux de données exportés des systèmes opérationnels e - Daniel Manage arbitre en lançant au préalable une opération de recensement/documentation des sources de données
	1		a
	2		b +e
	3		a+c
	4		d
	5		b+d+e
SF-BID 3			La vision SafeSecure24 vision Jean Desdonnet a pour intérêt : a - De permettre plus rapidement la consolidation des données b - De faciliter la parallélisation des opérations de collecte des données et celles de mise en qualité c - De faciliter la prise en main par les futur utilisateurs et développeurs d - De constituer un référentiel pour les CRM et ERP e - De raccourcir l'opération de recensement/documentation des sources de données dans laquelle Paul Desdonnet doit être impliqué Quelles sont celles que vous recommandez :
	1		a+b
	2		b
	3		c+d+e
	4		b+d
	5		d+e
SF-BID 4			Il existe actuellement de nombreux traitements statistiques locaux basés sur Excel. La mise en place de SafeSecure24 Solution Data va t-il poser un problème majeur dans le maintien de ces traitements : a - Les statistiques Excel vont pouvoir continuer à exister b - Il faudra envisager de les alimenter avec les sources issues de la SS24 Solution Data c - La bascule des données de HDFS vers Excel est complexe, il faut également ré-écrire les statistiques Excel d - Les calculs effectués par les traitements Excel ne peuvent pas être intégrés à une nouvelle plateforme HDFS e - Devant le constat qu'il existe beaucoup de calculs Excel, il faut mieux renoncer à utiliser HDFS
	1		a
	2		a+b
	3		c
	4		d
	5		a+e
SF-BID 5			Le cœur des données dans le Big data peut être le système de fichier mais également les bases de données de différentes natures : Colonne, NoSQL. Parmi ces choix : a - Utiliser principalement le système de fichier qui est compatible avec les sources non HDFS et facilite le travail avec les outils statistiques b - Utiliser dans tous les cas une base NoSQL qui permet tous les types de traitement c - Utiliser Cassandra qui est une base de données relationelle d - Utiliser selon les usages NoSQL et Base de données colonne e - Utiliser lorsque nécessaire des données en fichier
	1		a
	2		a+b

	3		c
	4		d
	5		d + e

SF-BID 6		**La règle des 3 V ne présente que des avantages pour traiter de la volumétrie, de variété des sources et de Vélocité de traitement. On peut donc dire :** **a - les 3V permettrons de faciliter le Data Mining** **b - La garantie de bonne fin des transaction est assurée** **c - Il faudra toutefois réduire le paramètre de réplication des données de Yarn à 1 pour obtenir des performances correctes** **d - La règle des 3V est incompatible avec la constitution d'un Data Warehouse sur l'environnement HDFS**	
	1		a
	2		b
	3		a + c
	4		d
	5		a + d

SF-BID 7		**La vision SafeSecure24 Solution Data de Paul Décalté facilite la constitution d'un pôle de données mixtes ERP/CRM avec référentiels commun ?** **a - Oui car les traitements peuvent être développés dans l'environnement actuel bien maitrisé par la DSI** **b - Pas nécessairement, car les outils de traitements nécessaires à créer ces référentiels sont plus simples en Big Data** **c - Les questions de rapidité et de paralélisation sont primordiales et donc militent pour le Big Data** **d - On peut dire que les 3V ont tout leur usage pour mixter des sources différentes comme ERP et CRM** **e - Le Big Data n'est pas fondalement fait pour ce type de traitements**	
	1		a
	2		b
	3		c
	4		c + d
	5		a + e

SF-BID 8		**Les données de CRM et ERP sont portés par des modèles différents, faut-il constituer des bases de données séparées dans l'environnement HDFS pour simplifier les choses ?** **a - Oui car il n'existe pas d'identifiants communs dans les systèmes actuels** **b - Pas nécessairement, si les traitements d'intégration effectuent ce calcul de rapprochement par identifiants communs** **c - Le NoSQL permet de créer des modèles mixtes contenant des données de sources différentes** **d - Ca n'a aucun intêret car c'est au final les traitements de Data Mining qui vont faire les rapprochements** **e - Construire un modèle intégré de type DWH est franchement à envisager**	
	1		a
	2		b
	3		e
	4		b + c + e
	5		b + d

SF-BID 9		**Les traitements R devraient être sur la plateforme SS24 Data Solution. Les Data Scientists vont pouvoir faire des traitements :** **a - Uniquement sur les données transformées en fichiers** **b - Uniquement sur les bases de données Hbase, Cassandra de l'environnement HDFS** **c - Sur les fichiers et les bases de données** **d - R va pouvoir exploiter les capacités de traitement parallèle de HDFS** **e - On peut utiliser Python pour disposer des fonctions MapReduce**	
	1		a
	2		b
	3		c
	4		c + d
	5		c + d + e

SF-BID 10		**Pour la visualisation des données, il existe chez SafeSecure24 une solution SAS et une solution Tableau Software. Ces deux solutions vont-elles devoir être remises en question ? :** **a - SAS ne peut pas se connecter à Hadoop** **b - Tableau Software est la seule solution en dehors des librairies graphiques R pour visualiser des statistiques issues du Big Data** **c - Il existe de nombreuses solutions compatibles Hadoop car la plupart des éditeurs ont développé des connecteurs** **d - Il n'existe aucune solution native Hadoop de visualisation avancée de données**	
	1		a
	2		b
	3		c
	4		c + d
	5		a + d