Профиль: «Системы больших данных / Big Data Systems» **КОД - 173**

1. Дайте развернутые ответы (на английском языке):

Задание 1.1. What is a "dark data" and how they can be used for analysis? (15 баллов)

There are many definitions of dark data. You can use any, for example, are listed below.

1. Gartner defines dark data as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only. Storing and securing data typically incurs more expense (and sometimes greater risk) than value. (Gartner IT Glossary)

2. Dark data is a type of unstructured, untagged and untapped data that is found in data repositories and has not been analyzed or processed. It is similar to big data but differs in how it is mostly neglected by business and IT administrators in terms of its value.

Dark data is data that is found in log files and data archives stored within large enterprise class data storage locations. It includes all data objects and types that have yet to be analyzed for any business or competitive intelligence or aid in business decision making. Typically, dark data is complex to analyze and stored in locations where analysis is difficult. The overall process can be costly. It also can include data objects that have not been seized by the enterprise or data that are external to the organization, such as data stored by partners or customers. (https://www.techopedia.com/)

3. Dark data is data which is acquired through various computer network operations but not used in any manner to derive insights or for decision making. The ability of an organisation to collect data can exceed the throughput at which it can analyse the data. In some cases the organisation may not even be aware that the data is being collected. IBM estimate that roughly 90 percent of data generated by sensors and analog-to-digital conversions never get used.

In an industrial context, dark data can include information gathered by sensors and telematics.

Organizations retain dark data for a multitude of reasons, and it is estimated that most companies are only analyzing 1% of their data. Often it is stored for regulatory compliance and record keeping. Some organizations believe that dark data could be useful to them in the future, once they have acquired better analytic and business intelligence technology to process the information. Because storage is inexpensive, storing data is easy. However, storing and securing the data usually entails greater expenses (or even risk) than the potential return profit. (https://en.wikipedia.org/wiki/Dark_data)

Using of dark data can provide new insights that structured data assets currently in their possession may not reveal for examples in customer experience, business, and processes. New tools are based on advanced computer vision, pattern recognition, cognitive analytics, machine learning

There are some steps for using dark data:

- Find out what data is under company management and develop a plan what do with this data so that it delivers the value to the company.
- Find outside data sources that can enhance the value of data already have under enterprise management
- Quality assurance for data integrity and quality
- Develop proactive data management strategies for new technologies

Задание 1.2. How big data analytics affect the development of smart device? (15 баллов)

Big data analytics improves intelligence devices, allows you to integrate new services. This makes devices more functional, enhances their opportunities.

This allows you to create devices that better understand the consumer. They are easier to adapt to the interests of the consumer and, consequently, more demanded by consumers.

Smart devices based on big data allow to solve many tasks important for most areas:

- data collection and real-time response
- analysis of data coming from the sensors, combined with existing corporate or personal data to extract valuable knowledge
- using learned knowledge to refine and improve processes and applications

Combining big data analytics with smart devices is one of the ways to monetize data and expansion of market of decisions based on Analytics and Internet of things.

Particularly important impact on the development of industrial IoT. Implementation of analytics allows you to move from monitoring of products, customers, supply chains and objects to the decision-making and control actions, aimed, for example, at preventive maintenance

The decisions that can be based on knowledge gained from facts become less dependent on intuition and subjective experiences. So, costs can be reduced, processes can be streamlined and the quality of products and services can be increased.

2. Решите задачи:

Задание 2.1. (15 баллов)

Компания, занимающаяся маркетинговыми исследованиями, ведет базу данных для поддержания и автоматизации информационных потоков. Сведения о заказах включают следующие данные: номер, название заказа, даты старта и окончания выполнения заказа, перечень услуг, название контрагента, ФИО контактного лица, телефон, стоимость заказа, номер договора, дата договора, статус заказа.

Перечень оказываемых услуг содержит каталог из видов услуг и перечня маркетинговых исследований, их краткого описания. Для каждого заказа определяются сотрудники, которые будут выполнять работу, причем база данных содержит справочник со специализацией каждого сотрудника по видам оказываемых услуг и проводимым маркетинговым исследованиям.

Данные по контрагенту содержат: ИНН, вид контрагента, наименование, полное наименование, страну регистрации, КПП, ОГРН. По одному договору могут выполняться несколько заказов.

Требуется:

a) используя любую общепринятую нотацию, нарисовать схему базы данных, удовлетворяющую третьей нормальной форме, указать типы и направления связи;

b) сделать подробное описание таблиц с расшифровкой имен полей, указанием типов и свойств данных, ключевых полей;

c) используя операторы языка SQL, написать запросы для получения следующей информации:

- количество заказов, начатых выполняться в каждом месяце за текущий год. Результат представить в виде «месяц-количество заказов» по возрастанию месяцев;

- номера договоров, стоимость заказов по которым в прошлом.

Эталонное решение: работа 00119517

Задание 2.2. Найти значение параметра α, при котором функция $\alpha(1+x^2)^{-2}$ является плотностью распределения вероятностей некоторой случайной величины при $x \in (-\infty, +\infty)$. (15 баллов)

Решение:

Функция может характеризовать плотность распределения вероятностей некоторой непрерывной случайной величины, если:

1. $\alpha(1+x^2)^{-2} \geq 0$.

Данное неравенство выполняется для любых $x \in (-\infty, +\infty)$ при $\alpha \geq 0$.

2. $\int_{-\infty}^{+\infty} \alpha(1+x^2)^{-2}dx = 1$.

Вычисляя интеграл, получаем $\int_{-\infty}^{+\infty} \alpha(1+x^2)^{-2}dx = \alpha\pi/2 = 1$.

Следовательно, $\alpha = 2/\pi$

Ответ: $2/\pi$.

Задание 2.3. Найти значение параметра α, при котором функция $\alpha(1+x^2)^{-2}$ является плотностью распределения вероятностей некоторой случайной величины при $x \in (-\infty, +\infty)$. (10 баллов)

Решение:

Введем обозначение $\lambda = np$. Тогда формулу Бернулли можно представить в следующем виде:

$$P_n(m) = \frac{\lambda^m}{m!}\frac{n(n-1)\cdots\big(n-(m-1)\big)}{n^m}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^m}$$

Найдем предел последнего выражения при $n \to \infty$:

$$\frac{\lambda^m}{m!}\lim_{n\to\infty}\frac{n(n-1)\cdots\big(n-(m-1)\big)}{n^m}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^m}$$

Предел первой дроби:

$$\lim_{n\to\infty}\frac{n(n-1)\cdots\big(n-(m-1)\big)}{n^m} = 1$$

Предел знаменателя второй дроби:

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^m = 1$$

Предел числителя второй дроби:

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Таким образом,

$$\frac{\lambda^m}{m!}\lim_{n\to\infty}\frac{n(n-1)\cdots\left(n-(m-1)\right)}{n^m}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^m} = \frac{\lambda^m}{m!}e^{-\lambda}$$

Ответ: $P_m = [(np)^m/m!]e^{-np}$.

Задание 2.4. (15 баллов)

Решение:

Пронумеруем заданную систему контекстных подстановок P следующим образом:

S → xAB (1), S → x B (2), A → xAC (3),

A → xC (4), B → Dz (5), D→ y (6),

CD → CE (7) , CE → D (8), DE→ DC (9), Cz → Dzz (10).

Легко видеть, что среди всех выводимых цепочек наименьшей длиной обладает цепочка xyz, она выводится из S последовательным применением продукций (2), (5), (6).

Анализ структуры заданных подстановок показывает, что для порождения цепочек произвольно большой длины нужно сначала применить подстановку (1), а затем несколько раз подстановку (3).

Если подстановка (3) применена *n* раз, то получится цепочка вида

x^(n+1) C^n B, где x^(*n*+1) - это *n*+1 идущих подряд экземпляров буквы *x* (*n* не меньше 1), а C^n - это *n* идущих подряд экземпляров буквы *C*.

Буква B может быть заменена только подстановкой (5), тогда получится цепочка x^(n+1) C^n D z.

Размножение буквы z может быть достигнуто только последовательным применением подстановок (7), (8), (9), (10). Тогда подцепочка CD будет заменена подцепочкой Dzz, поэтому из цепочки x^(n+1) C^n Dz получится цепочка x^(n+1) C^(n-1) Dzz.

Легко видеть, что после выполнения n раз такого преобразования, мы получим цепочку x^(n+1) D^(n+1) z^(n+1), где n >= 1.

Применяя n+1 раз продукцию (6), мы каждый раз будем заменять очередное вхождение буквы D на букву y.

В итоге мы получим цепочку вида x^(n+1) y^(n+1) z^(n+1), где n >=1.

Ранее мы уже вывели цепочку xyz последовательным применением продукций (2), (5), (6).

Поэтому получаем ответ: из символа S всевозможными подстановками можно получить только цепочки из множества {x^n y^n z^n, где *n* >=1}.

Задание 2.5. (15 баллов) Плотность распределения вероятностей с носителем $(-\infty, +\infty)$ непрерывной случайной величины имеет следующий вид:

$$\rho(x) = A[1 - (1-q)x^2]^{1/(1-q)},$$

где $q < 3$ – параметр распределения, A – нормировочная константа.

Найти математическое ожидание случайной величины. (15 баллов)

Решение:

По определению математического ожидания абсолютно непрерывной случайной величины

$$M = \int_{-\infty}^{+\infty} x\rho(x)dx = A\int_{-\infty}^{+\infty} x[1 - (1-q)x^2]^{1/(1-q)}dx$$

Первообразная подынтегральной функции

$$\frac{(1 + (q-1)x^2)^{1+1/(1-q)}}{2q-4} + \text{const}$$

Найдем пределы

$$\lim_{x\to 0} \frac{(1 + (q-1)x^2)^{1+1/(1-q)}}{2q-4} = \frac{1}{2+2q}$$

$$\lim_{x\to \pm\infty} \frac{(1 + (q-1)x^2)^{1+1/(1-q)}}{2q-4} = 0$$

при $1 < q < 2$

Таким образом, $M = 0$ при $1 < q < 2$

Ответ: 0 при $1 < q < 2$.