بسمه تعالی

آزمون پایان‌ترم درس مبانی رایانش ابری
بهار ۱۴۰۱
زمان پاسخگویی: ۱۵۰ دقیقه

نام و نام خانوادگی:
شماره دانشجویی:

نکات:۱) گزینه درست در همین برگه امتحانی انتخاب شود. ۲) در حین امتحان سوال **پرسیده نشود**. (سرنوشت سوالات مبهم، بعد از آزمون مشخص می‌شود) ۳) نمره منفی نداریم (به تمامی سوالات پاسخ بدید).۴) پاسخ سوالات تشریحی را در پاسخنامه بنویسید.۶) هر سوال چندگزینه‌ای ۱ نمره از ۸۰ دارد. ۵) جدول کلمات در صحفه ۴ اورده شده است.

۱) به طور کلی بهره‌وری منابع در مراکز داده بالا نیست و بخش زیادی از انرژی در مراکز داده هدر می‌رود.
a. صحیح
b. غلط

۲) از کار افتادن یک DataNode یا disk failure در یک خوشه HDFS کل سیستم HDFS را با مشکل مواجه می‌کند (replication factor=3).
a. صحیح
b. غلط

۳) سیستم مدیریت منابع YARN، قابلیت مدیریت منابع را تنها برای یک چارچوب محاسباتی توزیع شده فراهم می‌کند. به عبارتی یک خوشه YARN را نمی‌توان همزمان برای MapReduce و Spark استفاده کرد.
a. صحیح
b. غلط

۴) کدام گزینه یک داده ساختارمند یا شبه-ساختارمند **نیست**؟
a. پایگاه داده رابطه‌ای
b. صفحات html
c. مجموعه‌ای از فایل‌های متنی
d. مجموعه‌ای از فایل‌های xml

۵) کدام گزینه از ویژگی‌های اصلی کلان داده **نیست**؟
a. حجم خیلی زیاد
b. سرعت تولید شدن بسیار بالا
c. تنوع بسیار زیاد
d. شبه-ساختارمند بودن

۶) کدام گزینه مثالی از یک فناوری که مقوله تحلیل کلان داده را ممکن می‌سازد، **نیست**؟
a. solid-state drives
b. tape drive
c. hard drives
d. cloud computing

۷) سیستم فایل توزیع شده HDFS برای سناریوهایی خوب عمل می‌کند که فایل‌های ...، ... نوشته می‌شوند و ... خوانده می‌شوند. (جای خالی از راست به چپ)
a. بسیار بزرگ، چندین بار، چندین بار
b. بسیار بزرگ، چندین بار، یک بار
c. بسیار بزرگ، یک بار، یک بار
d. بسیار بزرگ، یک بار، چندین بار
e. کوچک، چندین بار، چندین بار
f. کوچک، چندین بار، یک بار
g. کوچک، یک بار، یک بار
h. کوچک، یک بار، چندین بار

۸) به شکل کلی استفاده از هدوپ و HDFS برای سناریوهایی مناسب است که برنامه محاسباتی ... و داده‌هایی که پردازش می‌شوند .... هستند.
a. کوچک، کوچک
b. کوچک، بزرگ
c. بزرگ، کوچک
d. بزرگ، بزرگ

۹) کدام مولفه در HDFS وظیفه نگهداری مکان بلاک‌های داده را برعهده دارد؟
a. DataNode
b. Rack
c. NameNode
d. Client

۱۰) کدام گزینه مولفه‌ای از سیستم مدیریت منابع YARN است که وظیفه ان کاهش بار (load) مدیر منابع (Resource manager) برای هماهنگی و ایجاد منابع مورد نیاز برای انجام کارهای ارسالی است؟
a. Client
b. NodeManager
c. ApplicationMaster
d. Container engine

۱۱) در سیستم مدیریت منابع YARN کدام دسته از وظایف از مزیت محلی بودن داده می‌توانند بهره‌مند شوند؟
a. فقط map task
b. فقط reduce task
c. هر دو map task و reduce task

۱۲) کدام گزینه یک partitioning معتبر به عنوان ورودی پردازه‌های reduce در نتیجه اجرای یک Hadoop job است؟ (ساختار عبارت‌ها به صورت key<values> است.)

a. p0={tehran<20,30>, tabriz<15, 20>, shiraz<35,40,30>},
p1={ahvaz<45,40>,tehran<15>},
p2={mashhad<25,30,35>, esfahan<30,35>}

b. p0={tehran<20,30>, tabriz<15, 20>, shiraz<35,40,30>},
p1={ahvaz<45,40>, mashhad<25,30,35>,
esfahan<30,35>, shiraz<40>}
p2={zahedan<35,40>, tabriz<5,15,10>}

c. p0={tehran<20,30>, tabriz<15, 20>, shiraz<35,40,30>},
p1={ahvaz<45,40>},
p2={mashhad<25,30,35>, esfahan<30,35>}

d. p0={tehran<20,30>, tabriz<15, 20>, shiraz<35,40,30>,
mashhad<25>},
p1={ahvaz<45,40>},
p2={mashhad<25,30,35>, esfahan<30,35>}

۱۳) با توجه به متن درس، کدام گزینه در مورد چارچوب‌های محاسباتی مبتنی بر دیسک صحیح **نیست**؟

a. این چارچوب‌ها نتایج میانی را در دیسک می‌نویسند.
b. داده برای هر پرس‌وجو از دیسک خوانده می‌شود.
c. بازگشت از شکست در این چارچوب‌ها خیلی چالش برانگیز است.
d. بیشتر برای بارهای کاری Extract-Transform-Load کارائی بهتری دارند.
e. مثالی از این چارچوب‌ها Hadoop MapReduce است.

۱۴) با توجه به متن درس، کدام گزینه در مورد چارچوب‌های محاسباتی مبتنی بر حافظه صحیح **نیست**؟

a. برای بارهای کاری تکرارشونده مناسب هستند.
b. همانند چارچوب‌های مبتنی بر دیسک به حجم در دسترس و ویژگی‌های حافظه اصلی حساس نیستند.
c. مثالی از این چارچوب‌ها، Apache Spark است.
d. بازگشت از شکست در این چارچوب‌ها چالش برانگیز است.

۱۵) کدام گزینه شی اصلی یا هسته (core object) در سیستم Spark است؟

a. Log
b. Map task
c. Reduce task
d. RDD
e. Lineage

۱۶) به شکل کلی خروجی یک Transformation و یک Action در Spark به ترتیب... و ... است.

a. یک RDD و یک RDD
b. یک RDD و یک مقدار
c. یک مقدار و یک RDD
d. یک مقدار و یک مقدار

۱۷) امکاناتی که OpenStack فراهم می‌کند نزدیکتر به کدام گزینه است؟

a. Infrastructure as a service
b. Platform as a service
c. Software as a service
d. Function as a service

۱۸) کدام گزینه از ویژگی‌های OpenStack **نیست**؟

a. داشتن معماری پیمانه‌ای
b. متشکل شدن از یک مجموعه خدمات هسته و بسیاری خدمات دیگر
c. متن باز بودن
d. استفاده از چارچوب احراز هویت واحد برای همه خدمات
e. تسهیل در انجام Scale up

۱۹) کدام مولفه از OpenStack مسئول ایجاد ماشین‌های مجازی است؟

a. KeyStone
b. Cinder
c. Neutron
d. Swift
e. Glance
f. Nova

۲۰) کدام مولفه از OpenStack چارچوبی را برای SDN فراهم می‌کند؟

a. KeyStone
b. Cinder
c. Neutron
d. Swift
e. Glance
f. Nova

۲۱) کدام مولفه از neutron مدل شبکه را مدیریت و اعمال می‌کند؟

a. L3-agent
b. Plugin-agent
c. DHCP-agent
d. Message queue
e. Neutron-server

۲۲) اگر دسترسی شبکه مابین ماشین‌های مجازی موجود در یک میزبان را لازم داشته باشیم اما ماشین‌های مجازی نیازی به دسترسی به خارج از میزبان را نداشته باشند، کدامه مولفه از neutron بایستی در هر compute node نصب و راه‌اندازی شود؟

a. L3-agent
b. Plugin-agent
c. L3-agent and plugin-agent
d. DHCP-agent
e. DHCP-agent and plugin agent
f. DHCP-agent and L3-agent
g. Neutron-server

۲۳) در یک خوشه OpenStack دستوری توسط nova-compute برای ایجاد یک ماشین مجازی به یک compute node ارسال شده است. این دستور بر روی کدام شبکه انتقال داده می‌شود؟

a. Guest network
b. Management network
c. External network
d. The Internet

۲۴) به شکل کلی با انتقال از مراکز داده سنتی به مراکز داده مدرن تعداد کاربردها ... و مقیاس (scale) آنها ... پیدا می‌کند. (جای خالی به ترتیب از راست به چپ)

a. افزایش، افزایش
b. افزایش، کاهش
c. کاهش، افزایش
d. کاهش، کاهش

۲۵) فرض کنید که کتابخانه‌ای در اختیار شما قرار داده شده است که ارتباط شما با یک سیستم تحت وب را فراهم می‌سازد. این کتابخانه از ادرس IP کارپذیرهای پاسخگوی درخواست اگاه است و برای هر درخواست کارپذیری را به شکل تصادفی انتخاب می‌کند. این سیستم تحت وب از کدام نوع است و اگر از این کتابخانه در برنامه خود استفاده کنید توزیع بار در چه سطحی انجام شده است؟

a. Cluster-based و Client
b. Cluster-based و DNS level
c. Cluster-based و Network level
d. Cluster-based و Dispatching device
e. Distributed و Client
f. Distributed و DNS level
g. Distributed و Network level
h. Distributed و Dispatching device

۲۶) توزیع کننده باری را در نظر بگیرید که بر اساس نام فایل درخواست کار توزیع بار را انجام می‌دهد و کارپذیر انتخاب شده خود مستقیما به client پاسخ را ارسال می‌کند. این روش مطابق با کدام گزینه است؟

a. سوئیچ لایه ۴ و معماری one-way
b. سوئیچ لایه ۴ و معماری two-way
c. سوئیچ لایه ۷ و معماری one-way
d. سوئیچ لایه ۷ و معماری two-way

۲۷) به طور کلی، الگوریتم‌های ایستای توزیع بار در مقایسه با الگوریتم‌های پویا، سرعت تصمیم‌گیری ... و کیفیت تصمیم ... دارند.

a. بالاتر، بالاتر
b. بالاتر، پایین‌تر
c. پایین‌تر، بالاتر
d. پایین‌تر، بالاتر

۲۸) با در نظر گرفتن فضای ابر عمومی، فراهم کنند ابر به کدام یک از داده زیر برای مدیریت مقوله تداخل عملکرد دسترسی **ندارد**؟

a. میزان استفاده از منابع مختلف در کارپذیر فیزیکی
b. میزان استفاده از منابع توسط هر کدام از ماشین‌های مجازی
c. کاربردهایی که داخل ماشین‌های مجازی اجرا می‌شوند.
d. شدت تداخل کارائی مابین ماشین‌های مجازی

۲۹) با توجه به انچه در کلاس درس بیان شد هنگامی که تداخل عملکرد برای LLC برای یک ماشین مجازی وجود دارد ... ، ... می‌یابد.

a. میزان استفاده از CPU، کاهش
b. میزان استفاده از CPU، افزایش
c. Disk wait time، کاهش
d. Disk wait time، افزایش

۳۰) در فرمول زیر، رابطه load با fg_load و bg _load چیست و کدام یک از درون ماشین مجازی قابل اندازه‌گیری **نیست**؟

$$T_{90} = c_0 + \frac{c_1}{(1-load)} + \frac{c_2}{(1-load)^2}$$

a. fg_load ،bg_load
b. bg_load ،bg_load
c. fg_load ،fg_load
d. bg_load ،fg_load
e. fg_load ،|bg_load-fg_load|
f. bg_load ،|bg_load-fg_load|
g. fg_load ،bg_load+fg_load
h. bg_load ،bg_load+fg_load

سوالات تشریحی (خوب فکر کنید و پاسخ بدید):

۱. عملگر Shuffle یک dataset (مجموعه‌ای از رکوردها) را می‌گیرد و به شکل تصادفی dataset را بهم می‌ریزد (re-order) (به عبارتی dataset خروجی دارای ترتیب تصادفی و متفاوتی از رکوردها است). از طرفی فرض کنید که تابع random(m) به شما داده شده است که یک عدد تصادفی در بازه [1 m] تولید می‌کند. شبه کد مپ-ردیوسی بنویسید که عملگر Shuffle را پیاده‌سازی کند. (**۵** نمره)

۲. فرض کنید که دو لیست به شما داده شده است. در لیست اول اطلاعات رای دهندگان وجود دارد:

(voter-id, name, age, zip)

در لیست دوم اطلاعات بیماری وجود دارد:

(zip, age, disease)

میخواهیم برای هر زوج یکتای age و zip، لیستی از نام‌ها و لیستی از بیماری‌های موجود در ان zip و با سن age را داشته باشیم. شبه کد مپ-ردیوسی بنویسید که این خروجی را برای ما بوجود اورد. به نکات زیر توجه کنید (**۷.۵** نمره):

- اگر امکان انجام این کار با یک map-reduce job وجود ندارد، می‌توانید از دو یا بیشتر map-reduce job کمک بگیرید به شرط انکه توضیح دهید ارتباط انها چیست و خروجی نهائی حاصل شود.
- اگر یک zip/age خاص در یک لیست ظاهر می‌شود اما در لیست دیگر وجود ندارد، می‌توانید در خروجی یک لیست خالی از اسامی یا یک لیست خالی از بیماری نشان دهید یا اینکه ان zip/age را به کلی از خروجی حذف کنید (به انتخاب خودتان).

۳. مولفه زمانبند در یک خوشه OpenStack برای استفاده از RamFilter پیکربندی شده است. همچنین پارامتر ram_allocation_ratio برابر با ۲ قرار داده شده است. از طرفی زمانبند از MetricWeigher استفاده می‌کند و پارامتراهای ان به شکل زیر مقداردهی شده‌اند (**۷.۵** نمره):

Weight_setting = available_CPU_cores = 0.5,

available_RAM = 0.3, available_disk = 0.2

در خواستی برای ایجاد یک ماشین مجازی با ۴ هسته CPU، ۱۶ گیگابایت حافظه اصلی و ۱۲۸ گیگابایت حافظه جانبی امده است. مشخص کنید کدام یک از کارپذیرهای زیر از فیلتر عبور می‌کنند و انها که عبور کرده‌اند به چه ترتیبی برای میزبانی این ماشین مجازی انتخاب می‌شوند. (نکته: برای مرحله وزن‌دهی بایستی داده‌ها را نرمال کنید و خط‌مشی معناداری را در نظر بگیرید).

	available_CPU_cores	available_RAM (MB)	available_disk (GB)
S1	2	10	8192
S2	4	7	512
S3	16	8	128
S4	16	16	64
S5	4	4	4096
S6	4	8	4096
S7	8	16	256

۴. چه روش‌هایی برای بهبود بهره‌وری منابع در مراکز داده وجود دارد؟ (حداقل **سه روش** را به قدر کفایت و با جزئیات کافی توضیح دهید، نیازی نیست دقیقا متن درس باشد و می‌توانید به دانسته‌ها و تجربه‌های خودتان رجوع کنید) (**۱۵** نمره)

۵. الف) با توجه به متن درس، **دو روش** توزیع بار لایه ۷ را تو ضیح دهید که مقوله کش شدن داده در حافظه اصلی کارپذیرها را در نظر می‌گیرند. توضیح شما بایستی نکات ضعف و قوت را شامل شود. (**۱۰** نمره)

ب) یکی از دو روش بالا را با یکی از روش‌های load sharing مقایسه کنید. (نیازی نیست دقیقا متن درس باشد و می‌توانید به دانسته‌ها و تجربه‌های خودتان رجوع کنید). (**۵** نمره)

جدول کلمات:

Resource utilization	بهره‌وری منابع
Framework	چارچوب
Distributed processing framework	چارچوب محاسباتی توزیع شده
Structured	ساختارمند
Semi-structured	شبه-ساختارمند
Load	بار
Task	وظیفه
Failure recovery	بازگشت از شکست
Iterative	تکرار شونده
Authentication	احراز هویت
Applications	کاربردها
Scale	مقیاس
Valid	معتبر
Network model	مدل شبکه
Load balancing	توزیع بار
Performance interference	تداخل عملکرد
Main memory	حافظه اصلی