يمكنالحوسبة، ITNT404

المحاضر:د. عمر أبو سعدة البريد الإلكتروني:

omar.abusaeeda@uot.edu.ly

Hadoop كأداة MapReduce

ماهو تقليل الخريطة؟

يمكنتعريف Map-Reduce على أنه**جزئين**:

-**عارضة**لكتابة البرامج التي يمكن إجراؤها بسهولة لمعالجة البيانات بالتوازي.

-**إطار العمل**الذي يقوم بتشغيل هذه البرامج بالتوازي، ويتعامل مع التفاصيل تلقائياً تقسيمالعمل والتوزيع والتزامن والتسامح مع الخطأ.

ال**نموذج**و ال**نطاق**العمل معاً لصنع برامج قابلة للتطوير، موزعةومتسامحة مع الخطأ.

أباتشيMapReduce

-العالم كله أصبح رقمياً

-إطار برمجي للمعالجة الموزعة لمجموعات البيانات الكبيرة

-يعتني الإطار**جدولة المهام**,**يراقب**لهم و**إعادة تنفيذ أي المهامالفاشلة.**

-يقوم بتقسيم مجموعة بيانات الإدخال إلى أجزاء مستقلة تتم معالجتها بطريقة متوازية تماماً.

-يقوم إطار عمل MapReduce بفرز مخرجات الخرائط، والتي يتم إدخالها بعد ذلك إلى الملف تقليلالمهام. عادةً، يتم تخزين كل من مدخلات ومخرجات المهمة في ملف نظام.

نموذجالبرمجة MapReduce

-**يتكون MapReduce من مرحلتين والابتكار الرئيسي الخاص به هو:**

-القدرة على إجراء استعلام على مجموعة بيانات وتقسيمها وتشغيلها بالتوازي على العديد من العقد.

-يحل مشكلة البيانات الكبيرة جداً (البيانات الكبيرة) التي لا يمكن احتواؤها على جهاز واحد

- الحوسبةالموزعة على العديد من الخوادم
- نموذجمعالجة الدفعات

-**مرحلة الخريطة**تتم معالجة بيانات الإدخال عنصراً تلو الآخر وتحويلها إلى وسيط

مجموعةالبيانات.

-**تقليل المرحلة**، يتم تقليل هذه النتائج الوسيطة إلى مجموعة بيانات مختصرة، وهي النتيجةالنهائية المرغوبة.

توزيعالبيانات

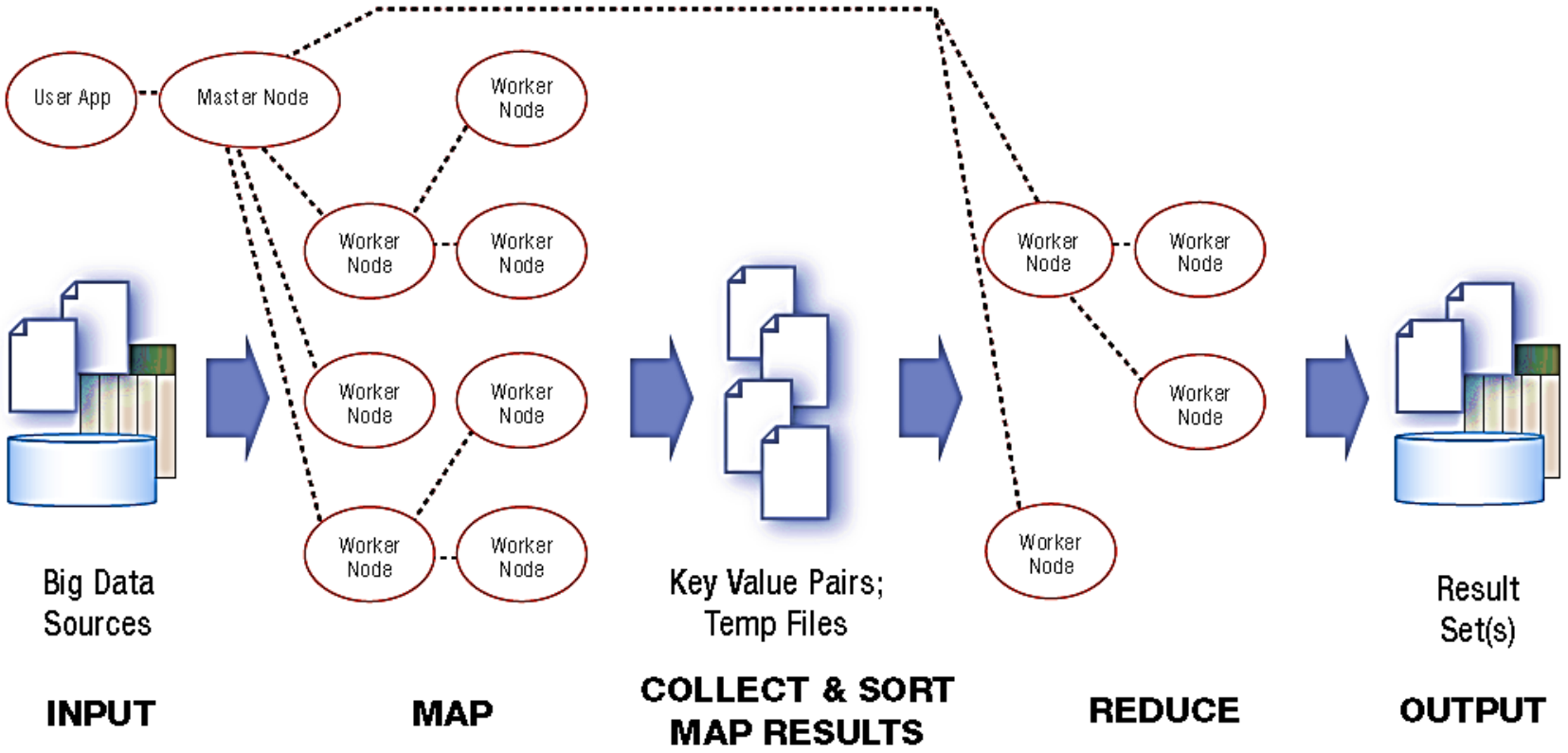
-في مجموعة MapReduce، يتم توزيع البيانات على جميع العقد في المجموعة أثناء تحميلها في.

-تقوم أنظمة الملفات الموزعة الأساسية (على سبيل المثال، GFS) بتقسيم ملفات البيانات الكبيرة إلى أجزاء تدارمن قبل العقد المختلفة في الكتلة

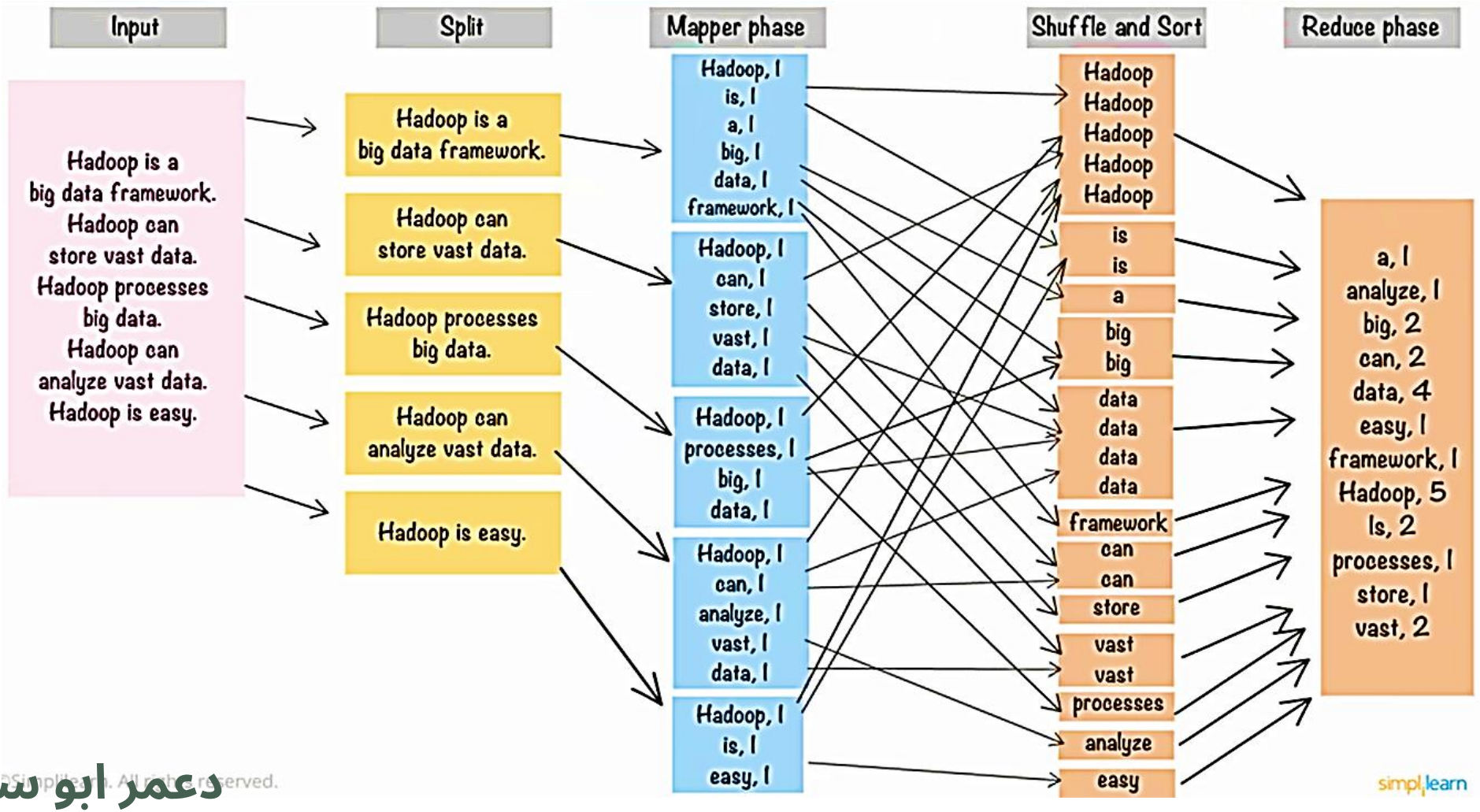
-على الرغم من أن أجزاء الملف يتم توزيعها عبر عدة أجهزة، إلا أنها تتشكل***مساحة اسم واحدة***

نظرةعامة على مستوى عالٍ جداً :MapReduce

-معالجة البيانات بطريقة موجهة نحو الدفُعات وقد تستغرق معالجتها دقائق أو ساعات (عادة).



نظرةعامة على مستوى عالٍ جداً :MapReduce



نظرةعامة على مستوى عالٍ جداً :MapReduce

-تحميل البيانات

- تسمىهذه العملية بشكل صحيح**يستخرج**,**تحول**,**حمولة**(**إيتل**) في مصطلحات تخزين البيانات.
- يجباستخراج البيانات من مصدرها، وتنظيمها لجعلها جاهزة للمعالجة، وتحميلها في ملف طبقةتخزين لـ MapReduce للعمل عليها.

MapReduce-

- ستقومهذه المرحلة باستعادة البيانات من التخزين،
- معالجتها(رسم خريطة، جمع نتائج الخريطة وفرزها، تقليلها)
- وإرجاعالنتائج إلى المخزن.

-استخراج النتيجة

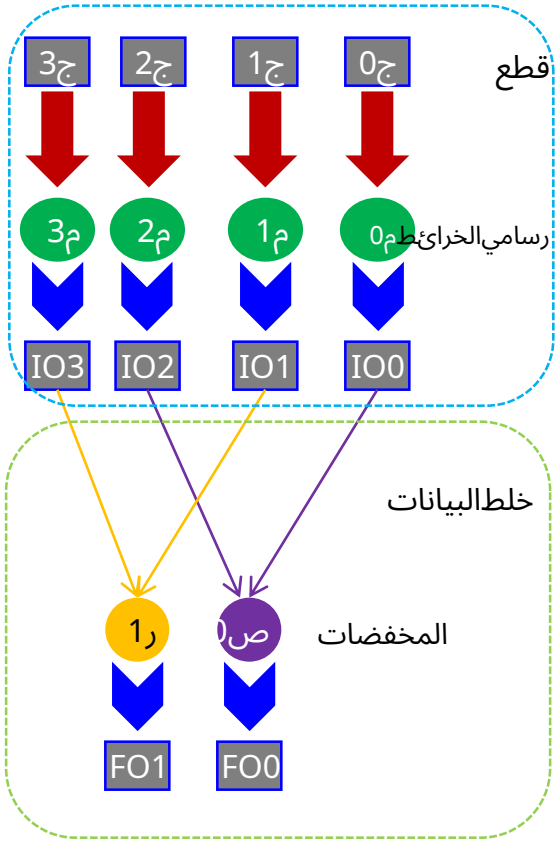
- بمجرداكتمال المعالجة، لكي تكون النتيجة مفيدة، يجب استرجاعها من المخزن و قدم.

نظرةشاملة :MapReduce

- فيMapReduce، تتم معالجة القطع بشكل منفصل عن طريق مهام تسمى ***مصممو الخرائط***
- تتمالإشارة إلى مخرجات مصممي الخرائط على أنها مخرجات وسيطة ويتمإحضارها إلى مجموعة ثانية من المهام تسمى)IOs(***المخففات***
- إنعملية جمع عمليات الإدخال والإخراج في مجموعة من المخففات معروفة مثل***عملية الخلط***
- تنتجالمخففات المخرجات النهائية (FOs)
- بشكلعام، يقوم MapReduce بتقسيم تدفق البيانات إلى مرحلتين،***مرحلة الخريطة*** و***تقليل المرحلة***

تقليل المرحلة

مرحلة الخريطة



نظرةعامة على مستوى عالٍ جداً :MapReduce

- MapReduce Programming Model
- Data type: key-value *records*

- Map function:

$$(K_{in}, V_{in}) \rightarrow list(K_{inter}, V_{inter})$$

- Reduce function:

$$(K_{inter}, list(V_{inter})) \rightarrow list(K_{out}, V_{out})$$

فوائدنموذج MapReduce

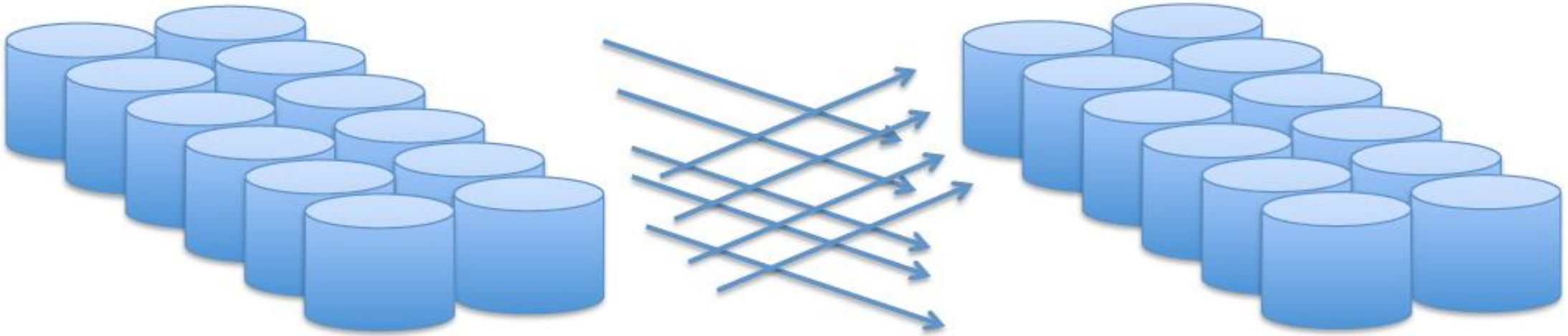
-من خلال توفير نموذج برمجة متوازي للبيانات، يمكن لـ MapReduce التحكم في المهمة التنفيذبطرق مفيدة:

- التقسيمالتلقائي للوظيفة إلى مهام
- الوضعالتلقائي للحساب بالقرب من البيانات
- موازنةالحمل التلقائي
- التعافيمن الفشل والمتعثرين

-يركز المستخدم على التطبيق، وليس على تعقيدات الحوسبة الموزعة (Implicit تماثل)

مثال:عدد الكلمات

-النمط الأساسي: سلاسل

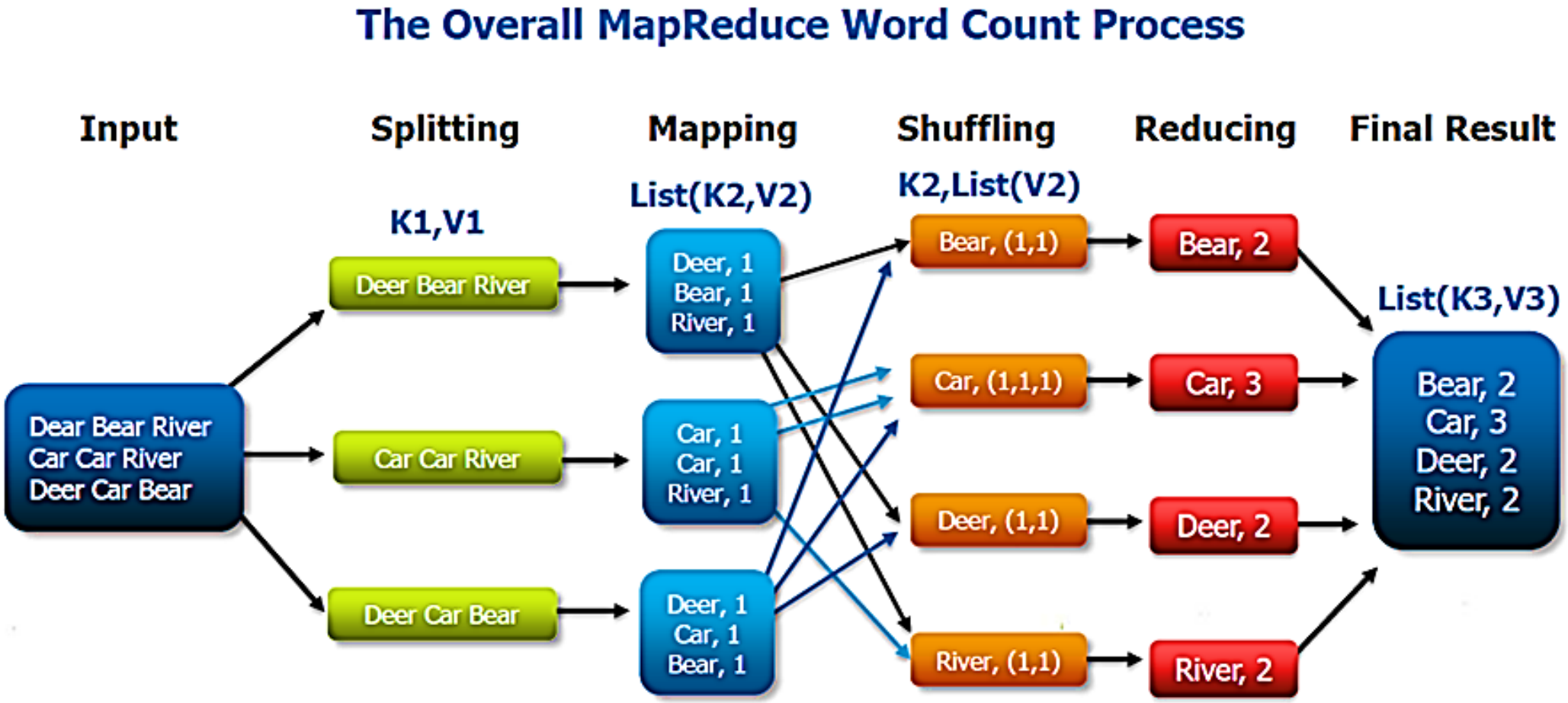


1.استخراج الكلمات من صفحاتالويب بالتوازي.

2.تجزئة الكلمات وفرزها.

3.العد بالتوازي.

عددالكلمات الشائعة



عددالكلمات الشائعة

```
خريطةباطلة (السلسلة i، خط السلسلة):
    للكلمةفي السطر:
        كلمةالطباعة، 1
```

عددالكلمات - وظيفة الخريطة

```
تقليلالفراغ (كلمة السلسلة، قائمة الأعداد الجزئية):
    المجموع= 0
    لـc في التهم الجزئية:
        المجموع+= ج
    طباعةالكلمة، المجموع
```

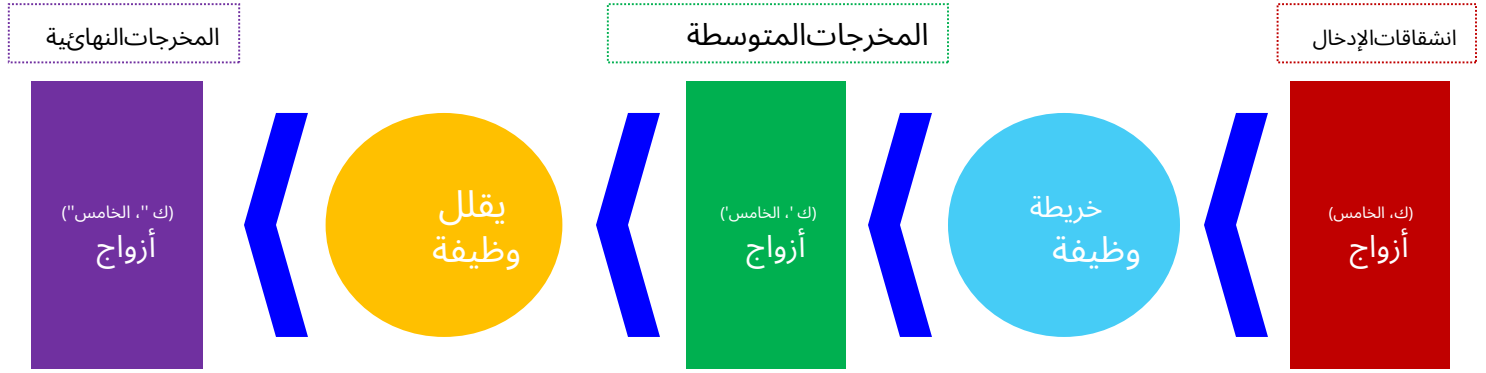
عددالكلمات - تقليل الوظيفة

المفاتيح والقيم

-يجب على المبرمج في MapReduce تحديد وظيفتين، ***وظيفة الخريطة*** و ال ***تقليل الوظيفة*** التي تنفذ Mapper وReducer في برنامج MapReduce.

-في MapReduce، يتم تنظيم عناصر البيانات دائماً كأزواج ذات قيمة رئيسية (على سبيل المثال، (K، V)).

-تقوم وظائف الخريطة والتقليل باستقبال وإصدار أزواج (K، V).



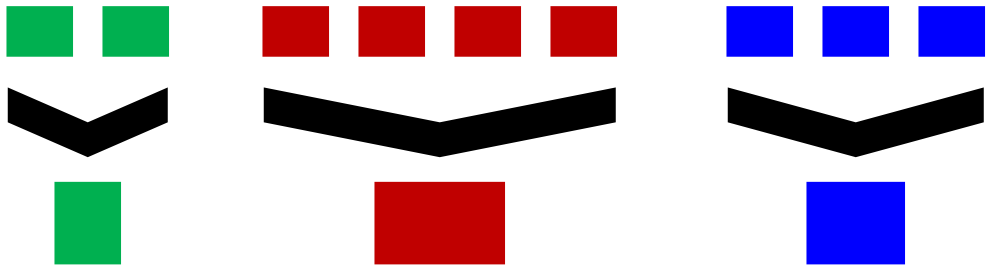
أقسام

-في MapReduce، لا يتم عادةً تقليل قيم المخرجات المتوسطة معاً

-يتم تقديم جميع القيم التي لها نفس المفتاح إلى مخفض واحد معاً

-وبشكل أكثر تحديداً، يتم تعيين مجموعة فرعية مختلفة من مساحة المفتاح المتوسطة لكل مخفض

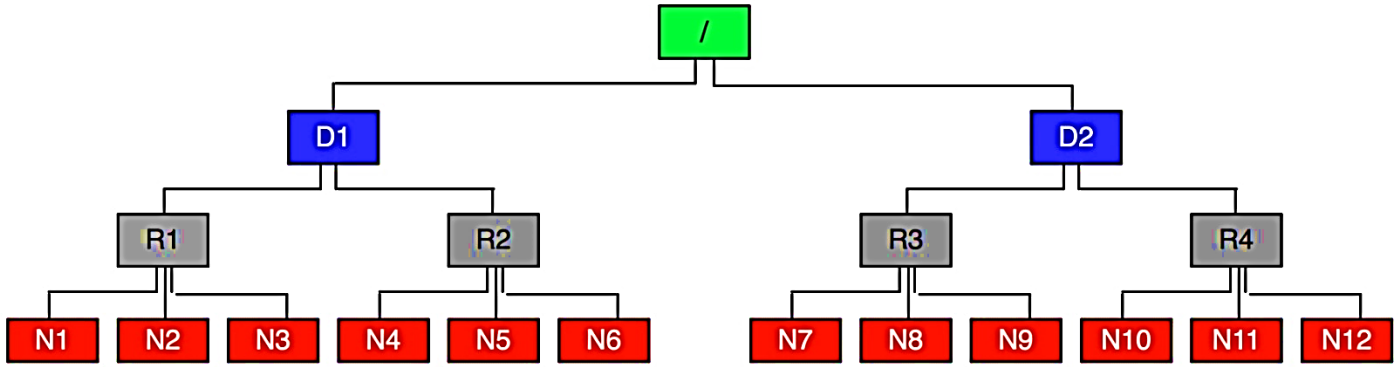
-تعُرف هذه المجموعات الفرعية بالأقسام



تمثل الألوان المختلفة مفاتيح مختلفة (من المحتمل) من مصممي الخرائط المختلفين

الأقسام هي المدخلات إلى المخفضات

طوبولوجياالشبكة في MapReduce



-يفترض MapReduce طوبولوجيا شبكة على شكل شجرة.

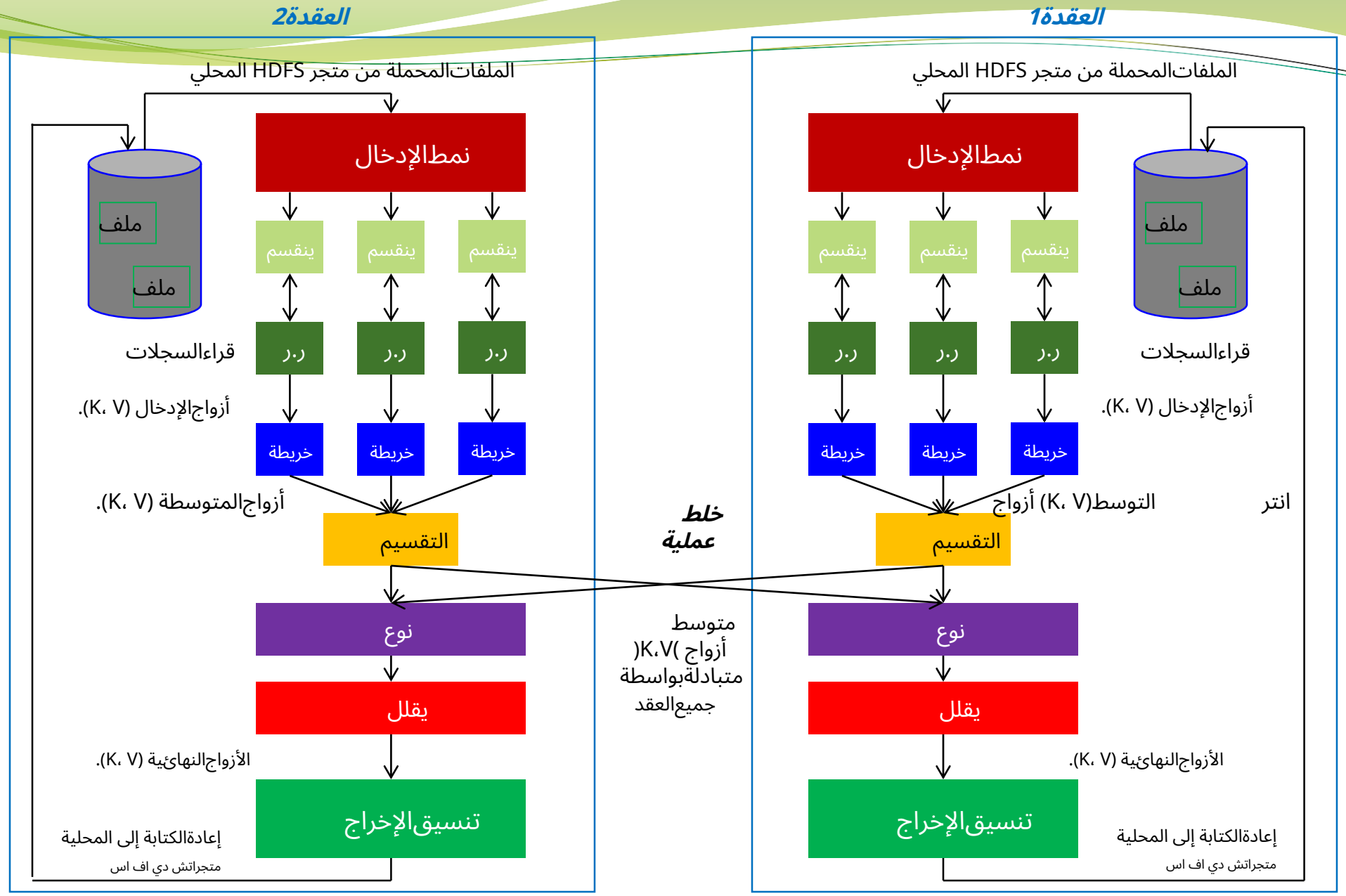
-تنتشر العقد على رفوف مختلفة متضمنة في مركز بيانات واحد أو أكثر.

-النقطة البارزة هي أن عرض النطاق الترددي بين عقدتين يعتمد على مواقعهما النسبية في طوبولوجيا الشبكة.

-على سبيل المثال، سيكون للعقد الموجودة على نفس الحامل نطاق ترددي أعلى فيما بينها

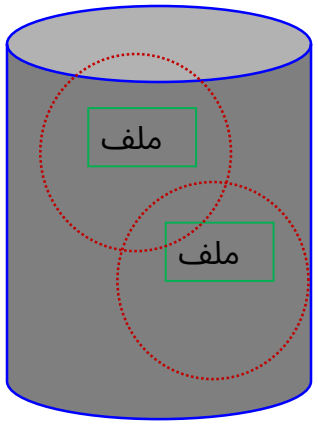
علىعكس العقد التي هي خارج الرف.

نظرةفاحصة: Hadoop MapReduce



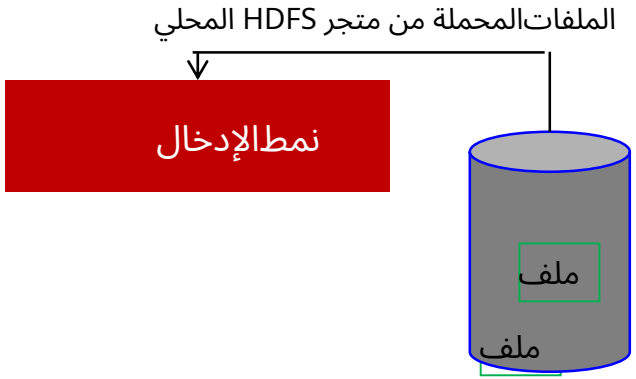
ملفاتالإدخال

- ***ملفاتالإدخال*** حيث البياناتل**MapReduce**يتم تخزين المهمة في البداية
- توجدملفات الإدخال عادةً في نظام ملفات موزع (مثل HDFS)
- تنسيقملفات الإدخال تعسفي
  - ملفاتالسجل المستندة إلى الخط
  - الملفاتالثنائية
  - سجلاتإدخال متعددة الأسطر
  - أوأي شيء آخر تماماً



نمطالإدخال

- تنسيقملفات الإدخال تعسفي
- يتمتحديد كيفية تقسيم ملفات الإدخال وقراءتها بواسطة Input_Format
- هيفئة تقوم بما يلي Input_Format:
  - تحديدالملفات التي يجب استخدامها للإدخال
  - يحددInputSplits التي تكسر الملف
- يوفر مصنعاً لكائنات Record_Reader التي تقرأ الملف



نوعتنسيق الإدخال

-يتم توفير العديد من تنسيقات الإدخال مع Hadoop:

نمطالإدخال	وصف	مفتاح	قيمة
تنسيقإدخال النص	التنسيقالافتراضي؛ يقرأأسطر الملفات النصية	إزاحةالبايت محتويات السطر	
KeyValueInputFormat	يوزعالخطوط إلى أزواج (K، V).	كلشيء إلىحرف علامة التبويبالأول	ماتبقى من الخط
SequenceFileInputFormat	خاصبـ Hadoop أداءعالي تنسيقثنائي	تعريفالمستخدم	تعريفالمستخدم

انشقاقاتالإدخال

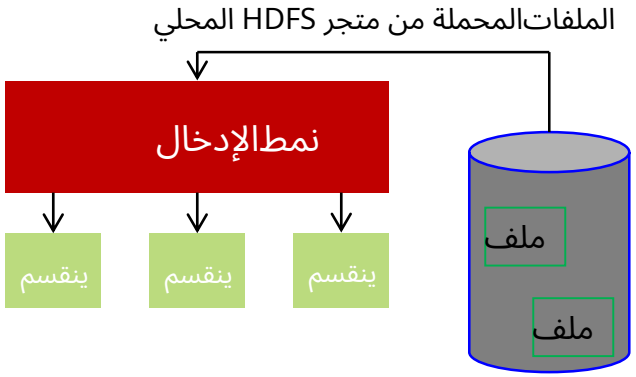
-ان***تقسيم المدخلات*** يصف وحدة العمل التي تشتمل على مهمة خريطة واحدة في برنامجMapReduce.

-افتراضياً، يقوم Input_Format بتقسيم الملف إلى أجزاء يصل حجمها إلى 64 ميجابايت.

-من خلال تقسيم الملف إلى أقسام، نسمح بتنفيذ العديد من مهام الخريطة تعملعلى ملف واحد بالتوازي.

-إذا كان الملف كبيراً جداً، فقد يؤدي ذلك إلى تحسين الأداء بشكل ملحوظ منخلال التوازي.

-تتوافق كل مهمة خريطة مع ***أعزب*** تقسيم المدخلات



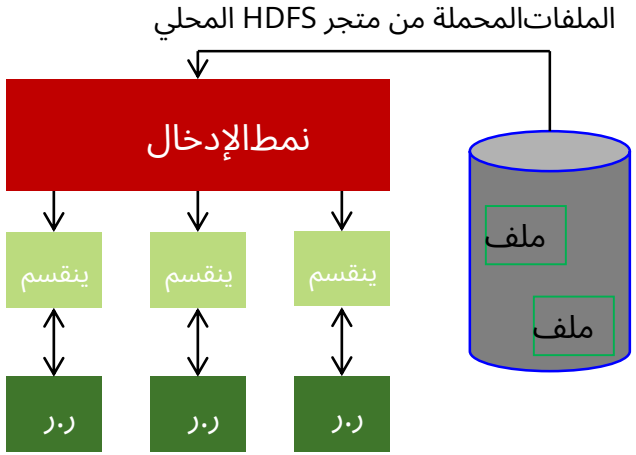
Record_Reader

-يحدد تقسيم الإدخال شريحة من العمل ولكنه لا يصف كيفية القيام بذلك للوصول إليه.

-ال***Record_Reader*** يقوم الفصل فعلياً بتحميل البيانات من مصدرها و يحولها إلى أزواج (K، V) مناسبة للقراءة بواسطة مصممي الخرائط.

-ال***Record_Reader*** يتم استدعاؤه بشكل متكرر على الإدخال حتى يتم استهلاك الانقسام بأكمله.

-كل دعوة من***Record_Reader*** يؤدي إلى استدعاء آخر لوظيفة الخريطة التي يحددها المبرمج.



مخططوالمخفض

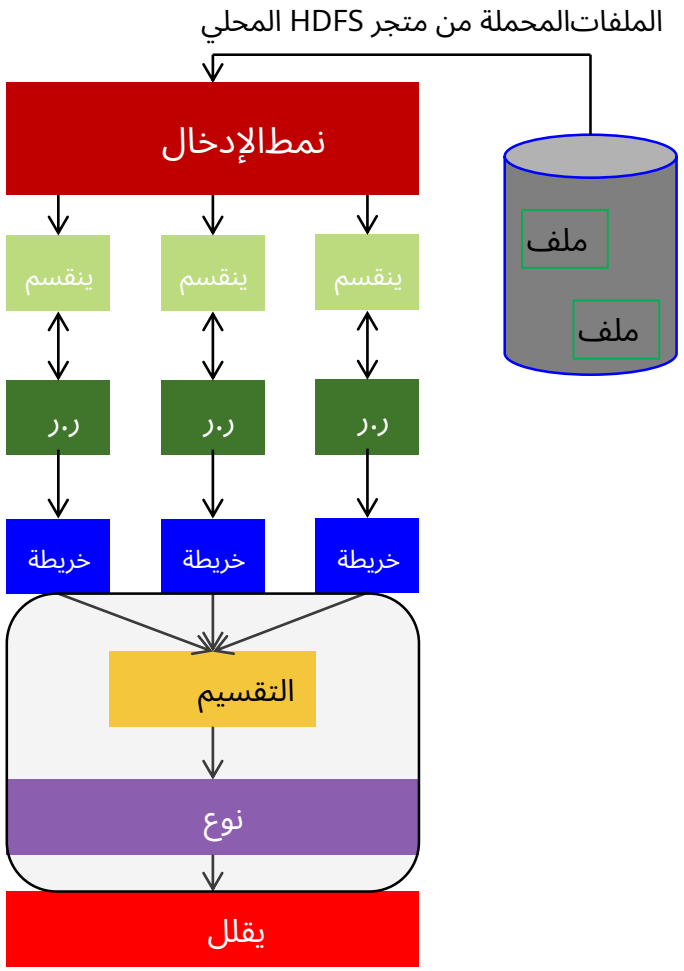
-ال***مصمم الخرائط*** ينفذ العمل المحدد من قبل المستخدم للمرحلة الأولىمن برنامج MapReduce

-يتم إنشاء مثيل جديد لـ Mapper لكل تقسيم

-ال***المخفض*** ينفذ العمل المحدد من قبل المستخدم للثانية مرحلةبرنامج MapReduce

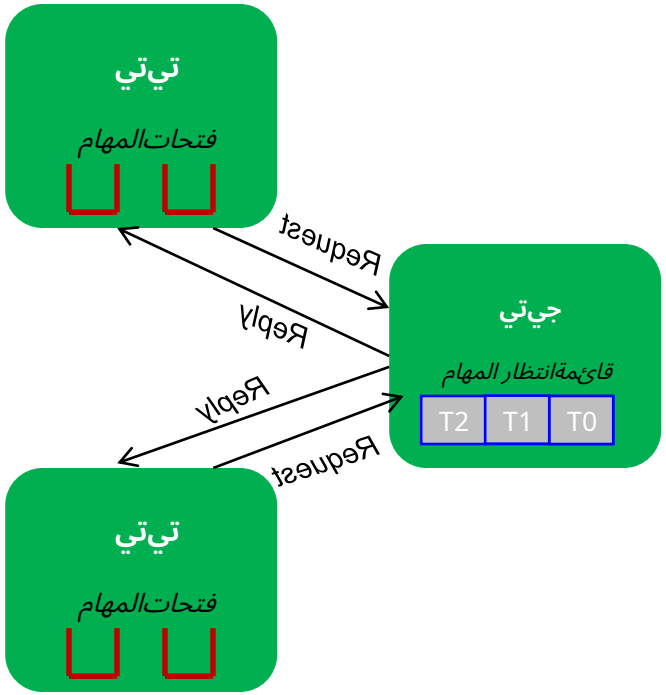
-يتم إنشاء مثيل جديد من المخفض لكل قسم

-لكل مفتاح في القسم المخصص للمخفض، فإن يتماستدعاء المخفض مرة واحدة.



جدولةالمهام في MapReduce

- يعتمد MapReduce ***ابنية السيد والعبد***
- يشار إلى العقدة الرئيسية في MapReduce باسم ***تعقب الوظيفة (JT)***
- تتم الإشارة إلى كل عقدة تابعة في MapReduce باسم ***تعقب المهام (TT)***
- يعتمد أ MapReduce ***جدولة السحب*** استراتيجية بدلا من دفعة واحدة
- على سبيل المثال، لا يقوم JT بدفع الخريطة وتقليل المهام إلى TTs بل يقوم TTs بسحبها عنطريق تقديم الطلبات ذات الصلة



Job_Tracker

-Job_Tracker في MapReduce هي الخدمة الروحية لإرسال وتتبع وظائف Hadoop.

-يقوم Job_Tracker بتنفيذ الإجراءات التالية في Hadoop:

- يقبلوظائف MapReduce من تطبيقات العميل
- يتحدثإلى Name_Node لتحديد موقع البيانات
- يحددموقع عقدة Task_Tracker المتاحة
- يرسلالعمل إلى عقدة Task_Tracker المختارة

Task_Tracker

-تقبل عقدة Task_Tracker الخريطة أو تقليل العمليات أو تبديلها عشوائياً من Job_Tracker.

-تم تكوينه بمجموعة من الفتحات، والتي تشير إلى عدد المهام التي يمكنه قبولها.

-يبحث Job_Tracker عن الفتحة المجانية لتعيين وظيفة.

-يقوم Task_Tracker بإعلام Job_Tracker بحالة نجاح المهمة.

-يرسل Task_Tracker أيضاً إشارات نبضات القلب إلى متتبع الوظائف للتأكد من توفرها، كما يقوم أيضاً بالإبلاغعن الرقم. من الفتحات المجانية المتاحة معها.

خريطةوتقليل جدولة المهام

- كلTT يرسل أ***رسالة نبضات القلب*** بشكل دوري إلى JT يتضمن طلباً لخريطة أو مهمة تقليل للتشغيل.

أنا.جدولة مهمة الخريطة:

- تلبيJT طلبات مهام الخريطة من خلال محاولة جدولة رسامي الخرائط في المنطقة

منانقسامات المدخلات الخاصة بهم (أي أنها تعتبر المنطقة).

ثانيا.تقليل جدولة المهام:

- ومعذلك، تقوم JT ببساطة بتعيين مهمة التخفيض التالية التي لم يتم تشغيلها بعد إلى TT الطالبة بغضالنظر عن موقع شبكة TT وتأثيرها الضمني على وقت التبديل العشوائي للمخفض (أي أنه لا يأخذفي الاعتبار المنطقة المحلية).

جدولةالوظائف في MapReduce

-في MapReduce، يتم تمثيل التطبيق على أنه***عمل.***

-تشمل الوظيفة خريطة متعددة وتقلل من المهام.

-يأتي MapReduce في Hadoop مع مجموعة مختارة من أدوات الجدولة:

- الافتراضيهو***جدولة FIFO*** الذي يقوم بجدولة الوظائف حسب ترتيب التقديم
- يوجدأيضاً برنامج جدولة متعدد المستخدمين يسمى***جدولة عادلة*** والتي تهدف إلى إعطاء كل استخدامحصة عادلة من سعة المجموعة مع مرور الوقت.

التسامحمع الخطأ في Hadoop

-يمكن لـ MapReduce توجيه المهام نحو إكمال ناجح حتى عند تشغيل المهام على مجموعة كبيرة حيث تزداد احتماليةالفشل

-الطريقة الأساسية التي يحقق بها MapReduce التسامح مع الخطأ هي من خلال***إعادة تشغيل المهام***

-إذا فشل TT في الاتصال بـ JT لفترة من الوقت (افتراضياً، دقيقة واحدة).

المعنيةقد تحطمت TT أن JT سوف تفترض ،)Hadoop

-إذا كانت المهمة لا تزال في مرحلة الخريطة، يطلب JT من TT آخر القيام بذلك***إعادة تنفيذ كافة مصممي الخرائط ذلك ركضسابقا في TT الفاشلة***

-إذا كانت المهمة في مرحلة التخفيض، فإن JT يطلب TT آخر***لإعادة تنفيذ كافة المخفضات ذلك كانتقيد التقدم في TT الفاشلة***

ماالذي يجعل MapReduce فريداً؟

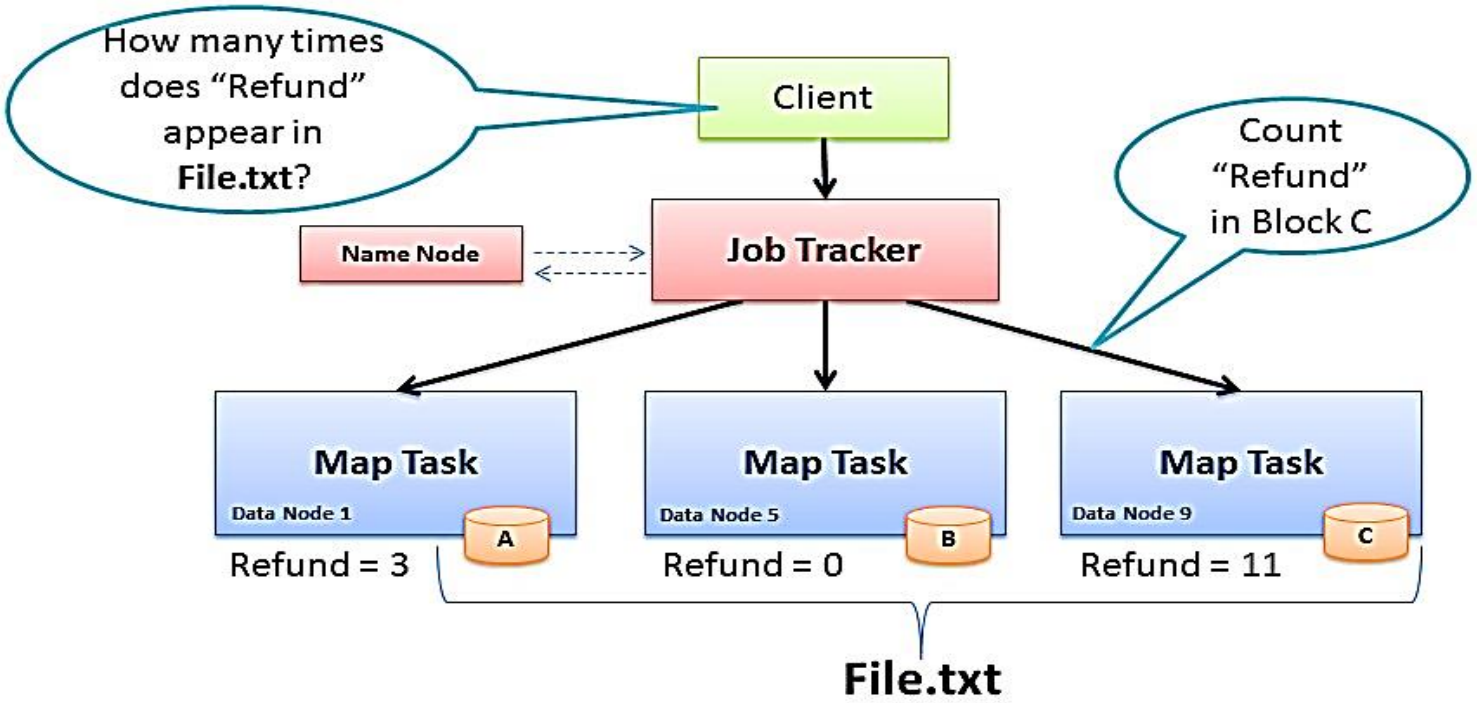
-يتميز MapReduce بما يلي:

1. نموذج البرمجة المبسط الذي يسمح للمستخدم بكتابة واختبار الأنظمة الموزعة بسرعة

2. التوزيع الفعال والتلقائي للبيانات وعبء العمل عبر الأجهزة

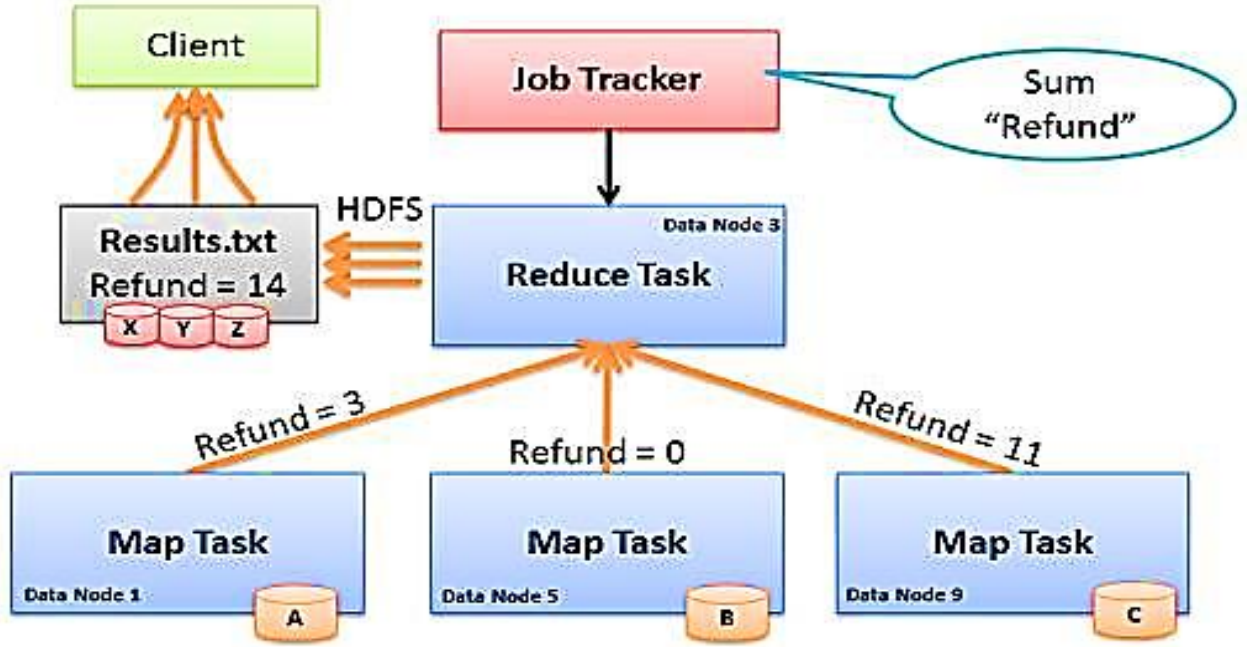
3. منحنى قابلية التوسع المسطح. على وجه التحديد، بعد كتابة برنامج MapReduce وتشغيله على 10 عقد،لا يلزم سوى القليل جداً من العمل، إن وجد، للقيام بذلك نفسالبرنامج يعمل على 1000 عقدة

مهمةالخريطة



- **Map:** "Run this computation on your local data"
- Job Tracker delivers Java code to Nodes with local data

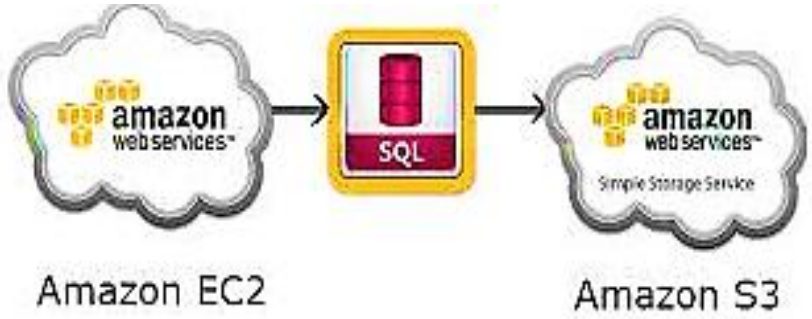
تقليل المهمة



- **Reduce:** "Run this computation across Map results"
- Map Tasks send output data to Reducer over the network
- Reduce Task data output written to and read from HDFS

AWS :في السحابة MapReduce

-يوفر واجهة قائمة على الويب وأدوات سطر أوامر لتشغيل وظائف Hadoop على Amazon ***EC2*** (سحابة الحوسبة المرنة من أمازون)



-البيانات المخزنة في الأمازون ***S3*** (خدمة التخزين البسيطة من أمازون)

-يراقب العمل ويغلق الآلات بعد الاستخدام

-رسوم إضافية صغيرة على رأس ***EC2*** التسعير

-ان ***EC2*** المثال يشبه جهاز كمبيوتر بعيد يعمل بنظام Windows أو Linux ويمكنك ذلك قمبتثبيت أي برنامج تريده، بما في ذلك خادم ويب يقوم بتشغيل كود PHP و خادمقاعدة البيانات.

-أمازون ***S3*** هي مجرد خدمة تخزين، تسُتخدم عادةً لتخزين الملفات الثنائية الكبيرة.

أمازون مطاطا MapReduce

مرونةMapReduce سير العمل

مرونةMapReduce سير العمل

مرونةMapReduce سير العمل

مرونةMapReduce سير العمل

المكونالمركزي لأمازون EMR

-المكون المركزي للأمازون**السجلات الطبية الإلكترونية**هل**تجَمَعّ**.

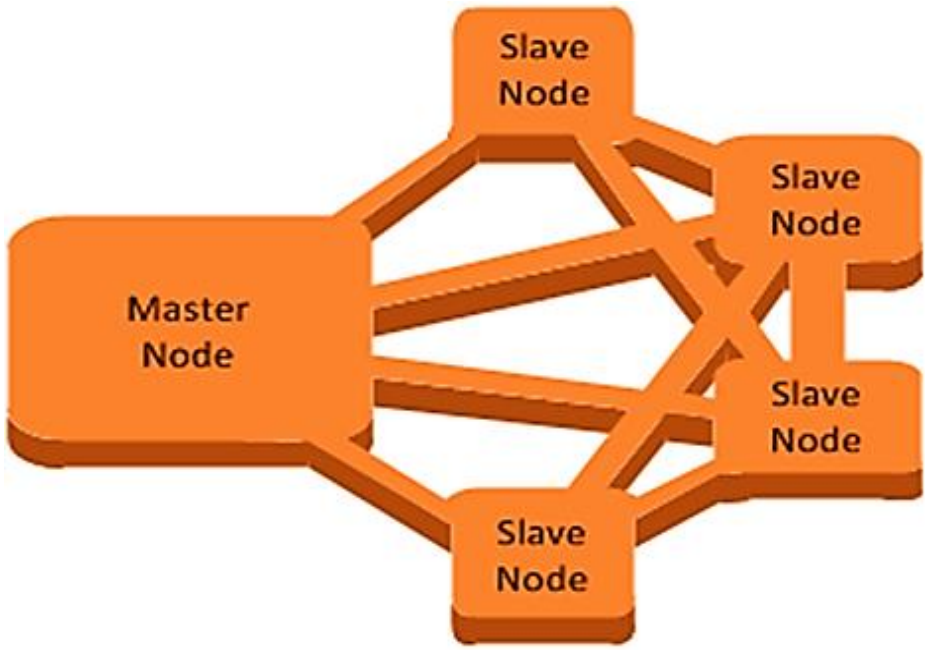
-**مجموعة**عبارة عن مجموعة من Amazon(Compute Cloud Amazon Elastic **EC2**) الحالات.

-**كل مثيل**في الكتلة يسمى**الأنود**.

-**كل عقدة**له دور داخل المجموعة، يشار إليه بنوع العقدة. يقوم EMR Amazon أيضاً بتثبيتات مختلفة

مكوناتالبرنامج على كل نوع عقدة، مع إعطاء كل منها

قمبعقد دور في تطبيق موزع مثل Apache Hadoop.



Cluster

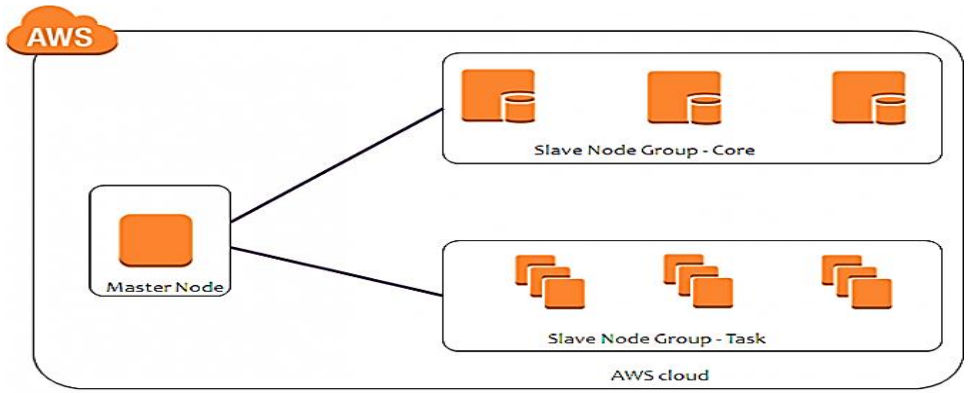
أنواعالعقد في Amazon EMR هي كما يلي:

- **العقدة الرئيسية:** عقدة تدير المجموعة عن طريق تشغيل مكونات برمجية لتنسيق توزيع البيانات والمهام بين العقد الأخرى- والتي يشار إليها مجتمعة باسم العقد التابعة -

  للمعالجة.تتتبع العقدة الرئيسية حالة المهام وتراقب صحة المجموعة.

- **العقدة الأساسية:** عقدة تابعة تحتوي على مكونات برمجية تقوم بتشغيل المهام وتخزين البيانات في Hadoop نظامالملفات الموزعة (HDFS) على مجموعتك.

- **عقدة المهمة:** عقدة تابعة تحتوي على مكونات برمجية تقوم بتشغيل المهام فقط. العقد المهمة هي خياري.



... شكراً لك ...