Hadoop

عادل قاضی خانی
هیئت علمی گروه کامپیوتر دانشگاه امام رضا (ع)

تعریف

- هدوپ بستری برای **ذخیره سازی داده های حجیم** بر روی **خوشه ای** [1] از کامپیوترهای ساده و **انجام عملیات روی داده ها** است

1-cluster

اجزای اصلی

- **دو جزء اساسی هدوپ عبارتند از:**

- بستر پردازش توزیع شده بنام **MapReduce**
- سیستم فایل توزیع شده بنام **HDFS**

1-cluster

عملکرد کلی

- هر برنامه ای که در هدوپ اجرا می شود **کار خود را بروی گره های (ماشین های) خوشه** تقسیم می کند.

- داده ها بر روی **HDFS** ذخیره شده و **MapReduce** پردازش را انجام می دهد.

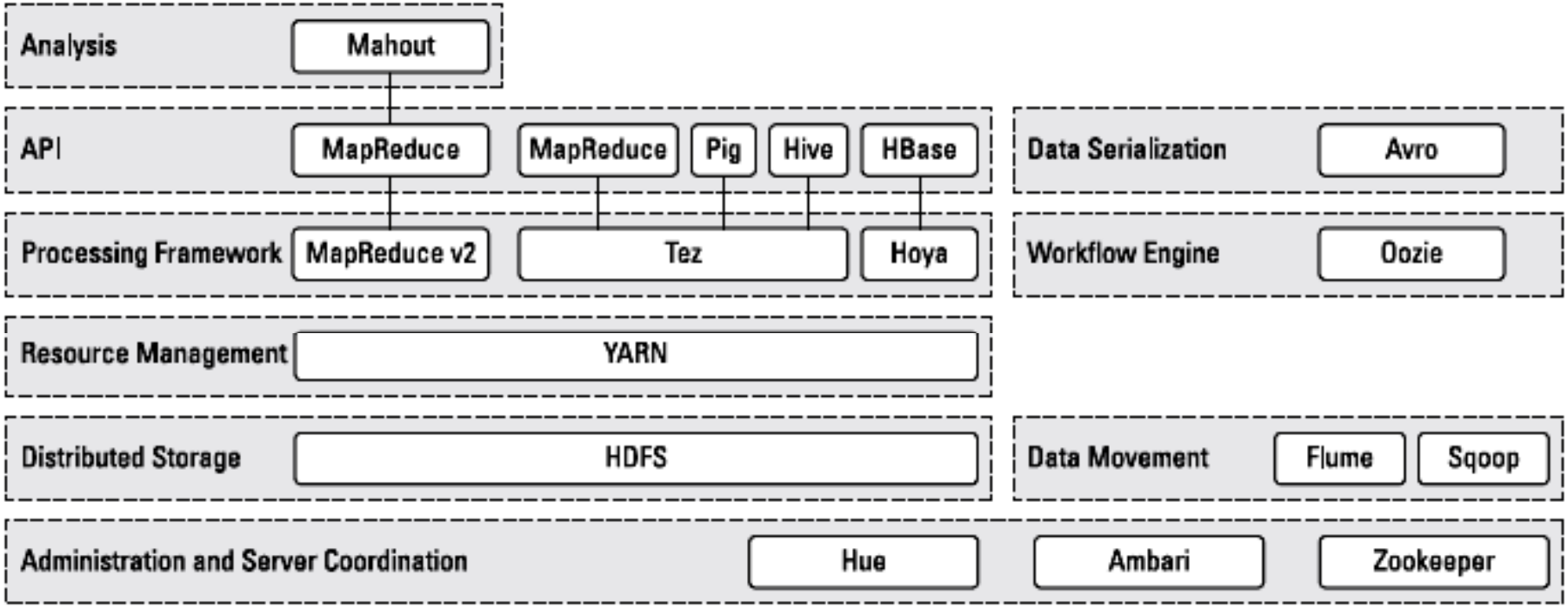
1-cluster

عملکرد کلی

- در خوشه هدوپ تعداد کمی **گره Master** و تعداد بیشتری **گره Slave** وجود دارد
- گره های **Master** فضای ذخیره سازی و سیستم پردازشی را **کنترل** می کنند
- گره های **Slave** که تمام داده ها روی آنها ذخیره شده و پردازشها در آنها **اجرا** می شود

1-cluster

ساختار کاملتر هدوپ



کاربرد هدوپ در داده های حجیم

- در ادامه چند کاربرد هدوپ برای تحلیل داده های حجیم بررسی می شود. کاربردها عبارتند از:

- تحلیل فایلهای log
- کمک به انباره داده
- تشخیص کلاه برداری
- تحلیل نظرات کاربران
- طبقه بندی تصاویر

تحلیل فایلهای log

- فایلهایی که **تمام عملیات کاربران** در سایتها در آنها ثبت می شود.
- **حجم** اطلاعات این فایلها **خیلی زیاد** است
- مثلاً وقتی برای خرید وارد یک سایت می شوید تمام کلیکهای شما ثبت می شود (**clickstream data**)

تحلیل فایلهای log

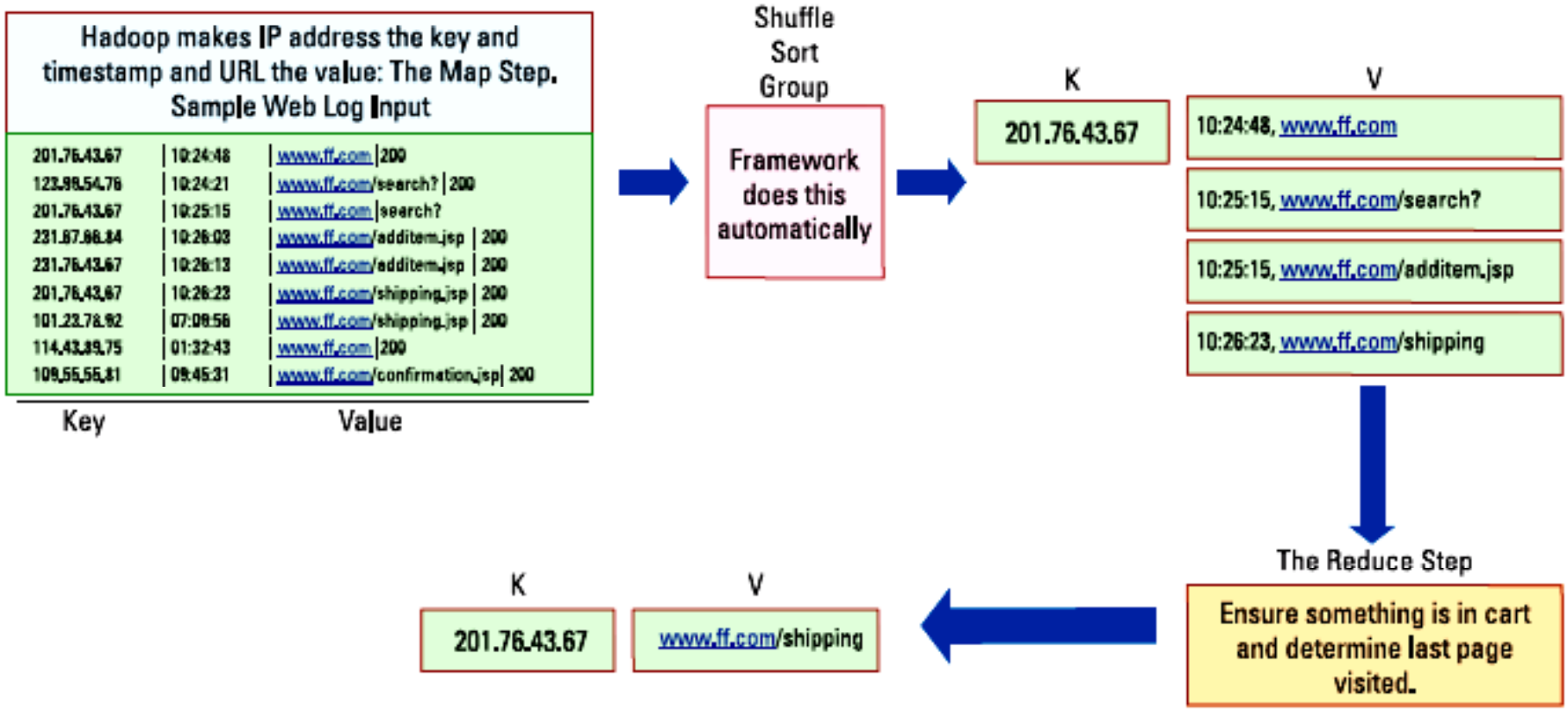
- تحلیل این فایلها در یک سایت فروش می تواند اطلاعات ارزشمندی در مورد **رفتارهای کاربران** در اختیار ما قرار دهد

- **مثلاً**: محصولاتی که نیمه کاره بدون خرید رها می شوند

تحلیل فایلهای log

- مثلاً می توانیم در این فایلها آدرس IP را بعنوان کلید داده ها در نظر بگیریم و مقدار آن برچسب زمانی و URL باشد

تحلیل فایلهای log

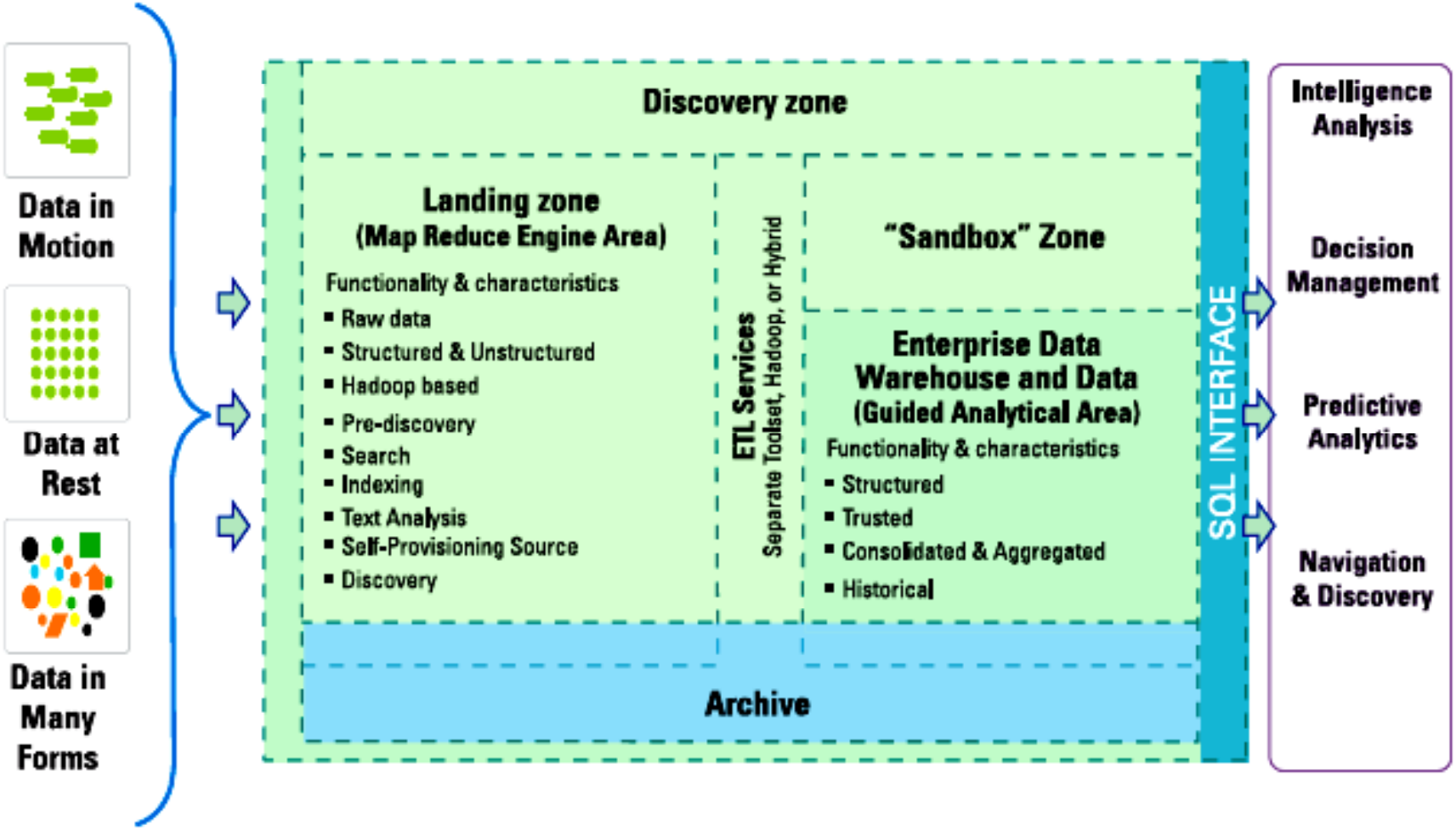


کمک به انباره داده

- **انباره داده** محلی برای ذخیره سازی داده ها از منابع مختلف سازمان به **منظور گزارش و تحلیل و تصمیم گیری**
- با افزایش حجم اطلاعات عملیات در انباره های داده نیز زمانبر خواهد بود
- **هدوپ** جایگزین انباره داده نیست بلکه **مکمل** آن است

Data warehouse

کمک به انباره داده



تشخیص کلاه برداری

- با استفاده از **روشهای داده کاوی** ما **بخشی از داده ها** را بعنوان نمونه انتخاب می کنیم و در ادامه با استفاده از یک الگوریتم داده کاوی تشخیص کلاه برداری انجام می شود

- در **هدوپ** می توانیم از **تمام داده ها** بهره بگیریم

Fraud detection

تحلیل نظرات کاربران

- **نظرات کاربران** در مورد مسائل گوناگون، بصورت متنی در رسانه های مختلف از قبیل **تالارهای گفتگو، وب لاگها، شبکه های اجتماعی** و ... منعکس می شود

- چون این **اطلاعات متنی** است برای پایگاه داده رابطه ای مناسب نیست

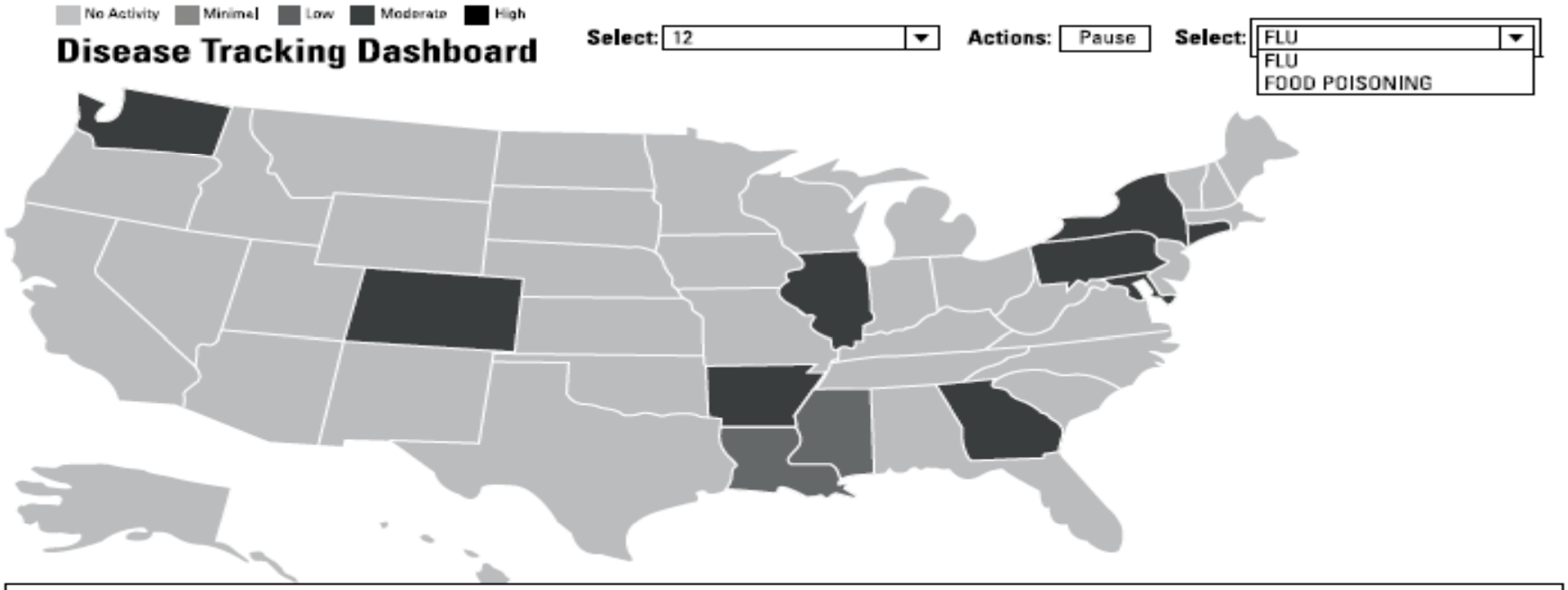
تحلیل نظرات کاربران

- **تحلیل متن** کار پیچیده ای است. کلمات معنایی متفاوت در مکانهای گوناگون دارند.
- زبانها به شکلهای گوناگون استفاده می شوند و در حال تکامل هستند
- یک روش برای تحلیل متن استفاده از **روشهای مبتنی بر گرامر** است

تحلیل نظرات کاربران

- **روش مناسب تحلیل متن داده کاوی** است
- مثلاً فرض کنید در حوزه سلامت وقوع بیماری آنفولانزا پیش بینی کنیم

تحلیل نظرات کاربران



Tweets

Sore throat has developed into full blown flu!

Though it was green apple two step - it's flu - both ends

Hands up for those who caught the flu bug as well

I think I got flu hate b in sick

I can feel flu coming on...hey that reminds me of a beer ad

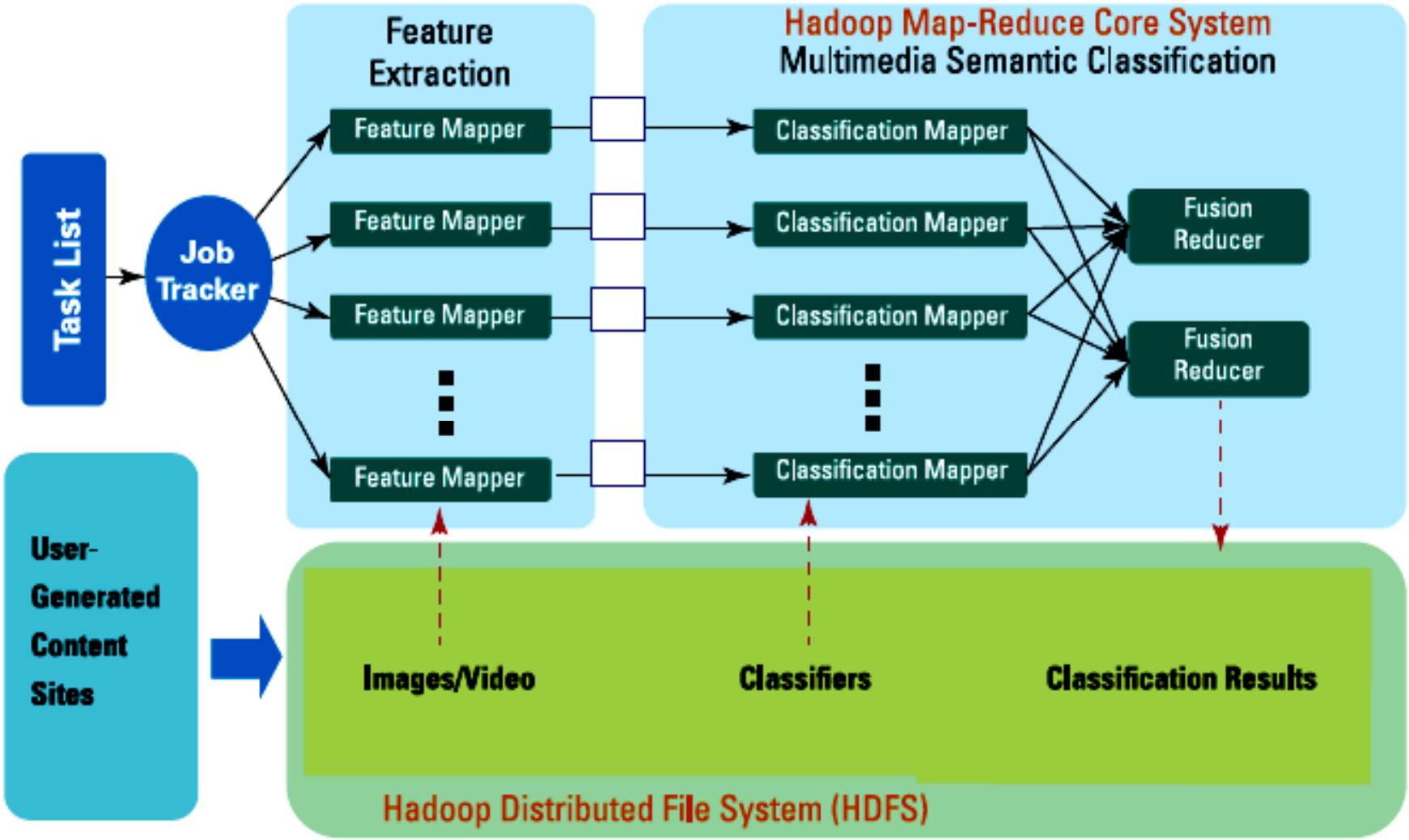
awful headache, shocking weather, flu, could today get worse #complaintover

Never want to get stomach flu again!! Dad next door is feeling worse than me

Thera flu for you

Morning peeps big day ahead back to work had 4 days off man flu

طبقه بندی تصاویر

- تصاویر در حوزه های مختلف تولید می شوند. مثلاً تصاویر چهره، عنبیه و ...
- **دو مرحله مهم** : استخراج ویژگی و طبقه بندی است

طبقه بندی تصاویر