

فهرست مطالب

Contents

[1 كليات تحقيق 3](file:///E:\projects\جلالي%20mapreduce\نهايي.docx#_Toc477514891)

[1-1 مقدمه 3](#_Toc477514892)

[1-2 معرفي الگوهاي مكرر 4](#_Toc477514893)

[1-3 كاهش فضاي جستوجو 6](#_Toc477514894)

[1-4 دادههاي عظيم 8](#_Toc477514895)

[1-4-1 معماري دادههاي عظيم 9](#_Toc477514896)

[1-4-2 MapReduce 10](#_Toc477514897)

[1-5 الگوهاي تكراري در دادههاي عظيم 11](#_Toc477514898)

[1-6 فرضيهها و اهداف تحقيق 13](#_Toc477514899)

[2 فصل دوم: پيشينه تحقيق 14](#_Toc477514900)

[2 پيشينه تحقيق 14](file:///E:\projects\جلالي%20mapreduce\نهايي.docx#_Toc477514901)

[2-1 مقدمه 15](#_Toc477514902)

[2-2 الگوريتمهاي استخراج الگوي مكرر تك متغيره 15](#_Toc477514903)

[2-3 استخراج الگوهاي مكرر غيرقطعي چندتايي 17](#_Toc477514904)

[2-4 استخراج الگوهاي مكرر غيرقطعي مجموعه اقلام 17](#_Toc477514905)

[2-4-1 روشهاي مبتني بر Apriori 18](#_Toc477514906)

[2-4-2 روشهاي مبتني بر FP-growth 20](#_Toc477514907)

[2-4-3 الگوريتمهاي مبتني بر H-Mine 23](#_Toc477514908)

[2-5 استفاده از MapReduce براي كاوش دادههاي عظيم 25](#_Toc477514909)

[2-6 استخراج الگوهاي تكراري با توجه به محدوديتهاي مورد نظر كاربران 26](#_Toc477514910)

[2-6-1 استخراج الگوهاي تكراري توسط مدل برنامهنويسي MapReduce 29](#_Toc477514911)

[3 فصل سوم معرفي روش پيشنهادي و زمينه هاي تحقيقاتي مرتبط 31](#_Toc477514912)

[3 معرفي روش پيشنهادي و زمينههاي مرتبط 31](file:///E:\projects\جلالي%20mapreduce\نهايي.docx#_Toc477514913)

[3-1 مقدمه 32](#_Toc477514914)

[3-2 استخراج الگوهاي مكرر از دادههاي غير قطعي 33](#_Toc477514915)

[3-3 تعريف داده هاي عظيم 35](#_Toc477514916)

[3-3-1 ويژگيهاي داده هاي عظيم 35](#_Toc477514917)

[3-3-2 ابزارهاي مفيد در تحليل دادههاي عظيم 37](#_Toc477514918)

[3-3-3 فايل سيستم توزيع شده هادوپ 39](#_Toc477514919)

[3-3-4 تابع نگاشت-كاهش يا MapReduce 43](#_Toc477514920)

[3-3-5 پايگاه داده NOSQL 45](#_Toc477514921)

[3-3-6 چالشهاي دادههاي عظيم 45](#_Toc477514922)

[3-4 استخراج الگوهاي مكرر تك متغيره از دادههاي عظيم 48](#_Toc477514923)

[3-5 استخراج الگوهاي مكرر چند متغيره از دادههاي عظيم 52](#_Toc477514924)

4 [پياده سازي و ارزيابي روش پيشنهادي 56](file:///E:\projects\جلالي%20mapreduce\نهايي.docx#_Toc477514925)

[4 فصل چهارم: پيادهسازي و ارزيابي روش پيشنهادي 56](#_Toc477514926)

[4-1 مقدمه 56](#_Toc477514927)

[4-2 مجموعههاي دادهاي 57](#_Toc477514928)

[4-2-1 مجموعه داده Conect4 57](#_Toc477514929)

[4-2-2 مجموعه داده mushroom 58](#_Toc477514930)

[4-2-3 مجموعه داده accident 58](#_Toc477514931)

[4-2-4 مجموعه داده IBM Quest Dataset Generator 59](#_Toc477514932)

[4-3 پيادهسازي روش پيشنهادي 60](#_Toc477514933)

[4-4 محيط ابر EC2 آمازون 61](#_Toc477514934)

[4-4-1 ويژگيهاي EC2 66](#_Toc477514935)

[4-4-2 برنامه نويسي مبتني بر MapReduce براي روش پيشنهادي 68](#_Toc477514936)

[4-4-3 آزمايش اول 73](#_Toc477514937)

[4-4-4 آزمايش دوم 74](#_Toc477514938)

[4-4-5 آزمايش سوم 77](#_Toc477514939)

[5 فصل پنجم: نتيجهگيري و پيشنهادات آينده 81](#_Toc477514940)

[نتيجهگيري و پيشنهادات آينده 81](file:///E:\projects\جلالي%20mapreduce\نهايي.docx#_Toc477514941)

[5-1 مقدمه 81](#_Toc477514942)

[5-2 خلاصه روش پيشنهادي 82](#_Toc477514943)

[5-3 نتيجهگيري 84](#_Toc477514944)

[5-4 پيشنهادات آينده 87](#_Toc477514945)

1

# كليات تحقيق

## مقدمه

پیشرفت­های به وجود آمده در دهه­های اخیر در زمینه جمع­آوری داده­ها و قابلیت ذخیره­سازی آنها باعث رویارویی با حجم بزرگی از اطلاعات در بسیاری از علوم شده است. داده کاوی کوششی جهت استخراج اطلاعات مفید از میان این داده هاست. طی سال­های گذشته در میان تکنیک­های داده کاوی توجه خاصی به الگوریتم­های کشف الگوهای مکرر[[1]](#footnote-2) شده است. الگوهای مکرر، مجموعه آیتم­ها، زیر ساخت­ها و یا توالی­هایی هستند که در یک مجموعه داده به تعدادی مساوی یا بیشتر از آنچه توسط کاربر به عنوان حد آستانه تعیین شده است تکرار شده باشند. مشکل اصلی در کاوش الگوهای مکرر پیدا کردن وابستگی آیتم ها در یک پایگاه داده است.

## معرفي الگوهاي مكرر

به طور كلي هدف داده­كاوي جست­وجو و استخراج اطلاعات ضمني و ناشناس و اطلاعات مفيد از داده­هاي اوليه است. روش­هاي مختلف داده­كاوي در برنامه­هاي كاربردي زندگي روزمره انسان استفاده مي­شوند به عنوان مثال گروه­بندي افراد بر اساس علايق آنها (خوشه­بندي[[2]](#footnote-3))، طبقه­بندي شركت­هاي بيمه بر اساس سن متقاضيان (كلاس­بندي[[3]](#footnote-4)) و شناسايي تراكنش­هاي اعتباري غير معقول (تشخيص ناهنجاري[[4]](#footnote-5)). در كنار خوشه­بندي، كلاس­بندي و تشخيص ناهنجاري، يافتن قوانين انجمني و الگوهاي تكراري نيز از اهميت ويژه­اي برخوردار هستند. زيرا اين دو روش مي­توانند به صاحبان و مديران فروشگاه­ها در يافتن الگوهاي جالب و پرطرفدار كه مشتريان بر اساس آن خريد مي­كنند، كمك بسزايي بكند.

فرض کنید مجموعه آیتم­های موجود در یک پایگاه داده تراکنشی مانند D باشد. مجموعه k آیتمی a که شامل k آیتم از مجموعه I است مکرر خواهد بود اگر و تنها اگر تعداد تراکنش­های موجود در D که شامل a هستند از کمتر نباشد. در این تعریف یک حد آستانه است که توسط کاربر تعیین شده است که پشتیبانی کمینه[[5]](#footnote-6) نامیده می­شود. تعداد تراکنش هاست و عدد بین صفر و یک است كه نمایانگر درصدی از تراکنش­هاست که شامل a هستند ]1[.

در دنیای واقعی همواره با دو دسته داده روبه رو هستیم ]2[:

1. داده­های دقیق[[6]](#footnote-7)، که با عدد صحیح و یا به مقدار دقیق بیان می­شوند.

2. داده­های غیر قطعی[[7]](#footnote-8)، با مقدار احتمالی مطرح می­شوند.

تحقیقات اخیر معمولاً از داده­های قطعی به سمت داده­های غیر قطعی سوق پیدا کرده است.

بسته به نوع داده غیر قطعی كه در کاوش الگوهای مکرر مطالعه می­شوند مي­توان آنها را به سه دسته تقسيم كرد ]3[:

1.داده­های غیر قطعی تک متغیره[[8]](#footnote-9)

2. داده­های غیر قطعی مجموعه اقلام[[9]](#footnote-10)

3. داده­های غیر قطعی چندتایی.[[10]](#footnote-11)

داده­های غیر قطعی تک متغیره به مواردی اطلاق می­شود که هر ویژگیدر هر تراکنش با یک بازه کیفی و یک تابع چگالی احتمال که احتمال هر مقدار را در یک بازده مشخص می­کند.

داده­های غیر قطعی مجموعه اقلام هر آیتم در یک تراکنش با یک احتمال در ارتباط است که مشخص کننده احتمال وجود آن آیتم در تراکنش است. و نوع سوم هر تراکنش به یک احتمال وابسته است. که نشان دهنده احتمال وجود آن تراکنش است

سه نوع داده ذکر شده از پایه با یکدیگر متفاوتند. داده­های غیر قطعی مجموعه اقلام را می­توان به صورت تراکنش­هایی شامل تعدادی از ویژگی­ها دانست که هر کدام با یک آیتم در ارتباط هستند، و محتوای هر ویژگی فقط نشان دهنده سطح احتمال وجود آن آیتم است. به طور واضح تر داده­های غیر واقعی تک متغیره شامل بازه­های پلکانی با احتمالات مربوط به آن می­باشد که مشخص کننده مقدار احتمال وجود آن است.

## كاهش فضاي جست­وجو

با توجه به افزايش چشمگير حجم داده­ها و گسترش چشمگير داده­هاي بزرگ[[11]](#footnote-12)، اخيراً اين بحث خيلي مورد علاقه محققين قرار گرفته است. در اين ميان آناليز و جستجو، در بين اين داده­ها مبحث مهمي به شمار مي­رود. با توجه به فضاي جستجوي وسيعي كه در اين نوع پايگاه داده­ها وجود دارد جستجو در بين آنها معمولاً فرآيندي زمان­بر و نادقيق است. فرض كنيد دو نفر يك كليد واژه يكسان را جستجو مي­كنند؛ آيا هميشه بايد نتيجه جستجوي آنها يكسان باشد؟ اگر هوشمندانه به اين قضيه نگاه كنيم جواب خير است زيرا افراد مختلف علايق و سليقه­هاي مختلف دارند و جوابي كه دريافت مي­كنند بايد بر اساس محدوده جستجوي آنها و علايقشان باشد. با رعايت اين موضوع دو دستاورد مفيد خواهيم داشت يكي اينكه نتايج جستجو دقيق­تر است و ديگر اينكه فضاي جستجو خيلي كاهش مي­يابد. در اين پژوهش بيشتر روي موضوع دوم يعني كاهش فضاي جستجو تاكيد مي­شود. اما چالش اصلي اين است كه چگونه علايق كاربران را به شكل يك الگوي مناسب تبديل كرد تا بتوان آن را در يك پايگاه داده بزرگ جستجو كرد.

از چالش­های عمده در کاوش الگوهای مکرر می­توان به برگرداندن مجموعه عظیمی از الگوها به عنوان پاسخ توسط الگوریتم­های کاوشی اشاره کرد. زمانیکه حد آستانه را کوچک انتخاب کنیم یا الگوهای مکرر طول بزرگی داشته باشد این مسئله بیشتر به چشم می­آید چرا که هر یک از الگوهای مکرر خود دارای زیر مجموعه­هایی هستند که آنها نیز مکررند. به همين دليل اين مسئله را در شاخه داده­هاي عظيم[[12]](#footnote-13) بررسي مي­كنيم.

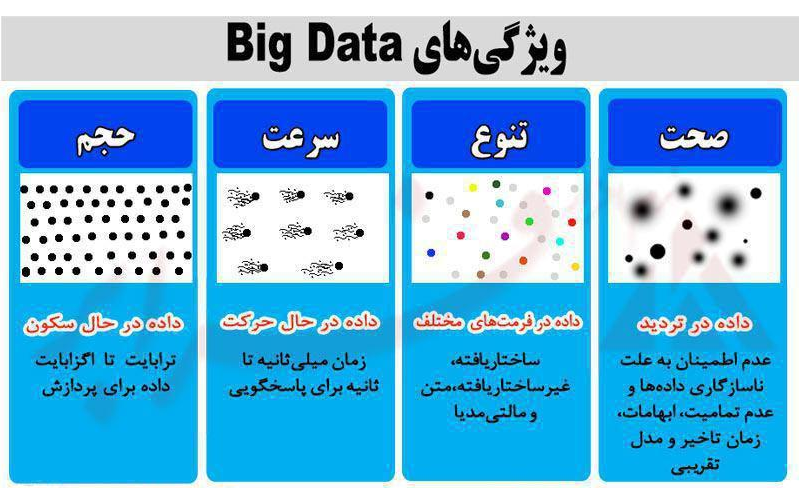
جهت غلبه بر این مشکل به معرفی دو مفهوم الگوی مکرر بسته[[13]](#footnote-14) و الگوی مکرر حداکثری[[14]](#footnote-15) می­پردازیم.

الگویP در مجموعه داده D یک الگوی مکرر بسته است و اگر P مکرر باشد و هیچ ابر الگویی مانند q از p وجود نداشته باشد که مجموعه پشتیبان آن دقیقاً برابر مجموعه پشتیبان P باشد ] 3 [.

الگوی P در مجموعه داده D یک الگوی مکرر حداکثری است اگر P مکرر باشد و هیچ الگویی از P مکرر نباشد.

## داده­هاي عظيم

با پيشرفت تكنولوژي حجم عظيمي از داده­ها در زمينه­هاي مختلف مانند جريان­هاي بانكي، مالي، بازاريابي، ارتباط از راه دور، بيولوژي، پزشكي، شبكه­هاي اجتماعي و غيره توسط برنامه­هاي كاربردي مختلف در حال شكل­گيري است. اين امر باعث پيدايش داده­هاي عظيم[[15]](#footnote-16) شده است ]18[. بحث داده­هاي عظيم مربوط به دستيابي به سرعت بالا، دقت بالا و تنوع زياد داده­ها در پردازش حجم زيادي از آن­ها است. جايي كه نرم­افزارهاي قديمي داده­كاوي قادر به پردازش روي اين حجم از داده­ها نيستند. بنابراين شيوه­هاي جديدي براي پردازش اين نوع داده­ها نياز است كه به آن كاوش داده­هاي عظيم[[16]](#footnote-17) مي­گويند.



شكل 1-1) ويژگي­هاي داده­هاي عظيم

### معماري داده­هاي عظيم

در سال 2000 شرکت سیزینت [[17]](#footnote-18)چهارچوب اشتراک فایل توزیع شده را مبتنی بر C++ برای ذخیره­سازی و جستجو در داده­ها توسعه داد که می­توانست داده­های ساختار یافته، نمیه ساخت یافته و غیرساخت یافته را در چندین سرور توزیع شده ذخیره و توزیع کند. در سال 2004 شرکت لکسیزنکسیز [[18]](#footnote-19)شرکت سیزینت و در سال 2008 شرکت چویزپوینت[[19]](#footnote-20) را به همراه پلترفرم­های پردازش موازی قدرتمند آنها را خرید و با ترکیب آنها سیستم HPCC را ارایه کرد که در سال 2011 تحت گواهی آپاچی 2 به صورت متن باز ارایه گردید. در حال حاضر HPCC و نیز Quantcast File System تنها سکوهایی هستند که به صورت عمومی در دسترس می باشند که قابلیت تحلیل هزاران ترابایت از داده­ها را دارا هستند.

گوگل نیز در سال 2004 سندی را برای پردازشی با نام MapReduce منتشر کرد که از همان معماری استفاده می­کرد. این چهارچوب نیز مدلی از پردازش موازی را ارایه می­کند که می­تواند پردازش­ها و جستجوها را تجزیه کرده و در نودهای مختلف مورد پردازش قرار دهد. این چهارچوب بسیار موفق بود و موجب شد تا دیگران نیز از همین الگوریتم استفاده کنند. بدین ترتیب پروژه ای بازمتن به نام Hadoop با به کارگیری این الگوریتم متولد شد. این چهار چوب از معماری (HDFS[[20]](#footnote-21)) بهره می­گیرد.

### MapReduce

براي مديريت داده­هاي عظيم محققين از يك مدل برنامه­نويسي سطح بالا به نام MapReduce استفاده مي­كنند كه قابليت پردازش بر روي حجم زيادي از داده­ها را به كمك محاسبات توزيع شده و موازي انجام مي­دهد. این فریم ورک داده‌های کلان را به زیرمجموعه‌ها تقسیم‌بندی می‌کند سپس آن‌ها را بر روی ماشین‌های مختلف تقسیم می‌کند سپس پردازش‌های جداگانه را در کنار هم می‌آورد و نتیجه نهایی را ارائه می‌دهد. این ایده از ایده معماری توزیع‌شده گرفته‌شده است. خرد کردن مشکل و سپس جمع‌آوری آن‌ها با نتایج .

## الگوهاي تكراري در داده­هاي عظيم

امروزه تعداد بیشتری از منابع داده، به ويژه پايگاه داده­هاي مربوط به داده­های بزرگ به شكل داده­هاي غیر قطعی هستند برای مثال: یک سنسور با حساسیت پائین برای جمع آوری و ثبت میزان آلودگی اتمسفر ممکن است فواصل کمی و غیر دقیقی از ذرات معلق را به جای مقادیر دقیق ثبت کند. بنابراین تابع چگالی احتمال به طور صریح یا ضمنی به بازه مشخص شده توسط حسگر اعمال می­شود تا احتمال وجود هر مقدار در آن بازه را تخمین بزند. به عنوان مثالی دیگر می­توان سیستم نظارت بر شبکه را نام برد که وظیفه دارد بازه­های پلکانی را برای تخمین ترافیک شبکه در هر ساعت جمع­آوری کند که تابع چگالی احتمال مشخص کننده احتمال وجود هر مقدار در بازه است.

با نظر به اینکه تعداد الگوهای فراوان در داده­های تک متغیره غیر قطعی به دلیل طول زیاد بازه­ها مقدار کوچک کمینه پشتیبان[[21]](#footnote-22) خیلی زیاد است، برای ارائه نتیجه به صورت مختصر و مفید به کاربر ما از مفهوم الگوی مکرر حداکثری داده­هی غیر قطعی تک متغیره استفاده می­کنیم.

از طرف دیگر در سالهای اخیراطلاعات سیستم­ها با سرعت زیادی تغییر می­کند و تراکنش­ها در پایگاه داده به سرعت و پیوسته به روز رسانی می­شوند. به عنوان مثال: بازار بورس، سیستم نظارت شبکه، مدیریت وب و ... به منظور استخراج کلیه الگوهای مکرر حداکثری از داده های غیر قطعی تک متغیره الگوریتم بایستی پس از بروز رسانی پایگاه داده مجددا روی کل پایگاه داده اجرا شود. چرا که ممکن است الگوهایی در پایگاه داده اولیه الگوی مکرر حداکثری نبوده­اند و با بروزرسانی تراکنش­ها تبدیل به الگوی مکرر حداکثری شده­اند یا الگوهایی که در پایگاه داده اولیه الگوی مکرر حداکثری بوده­اند اکنون جز الگوهای مکرر حداکثری نباشند. این در حالی است که قسمت اعظم پایگاه داده معمولا ثابت است و فقط قسمت کوچکی از پایگاه داده به روزرسانی می شود و اجرای مجدد الگوریتم هزینه های زیادی را در بر دارد. ما در این پژوهش با ارائه الگوریتمی سعی در استخراج کلیه الگوهای مکرر[[22]](#footnote-23) حداکثری داد های غیر قعطی تک متغیره در پایگاه داده های افزایشی داریم]13[.

بيشتر كارهاي انجام گرفته براي يافتن الگوهاي تكراري، روي مجموعه داده­هاي ساده كار كرده­اند اما در روش پيشنهادي يك مجموعه داده بزرگ مورد استفاده قرار مي­گيرد و الگوهاي تكراري روي اين مجموعه داده يافته مي­شود. اما مراحل روش پيشنهادي را مي­توان به سه بخش اصلي تقسيم كرد:

1- با توجه به اينكه در روش پيشنهادي از داده­هاي غير قطعي استفاده مي­شود در ابتدا سعي مي­شود الگوهاي تكراري روي پايگاه داده عظيم يافته شود.

2- محدوديت­ها و علايق كاربر توسط روش SAM كه تقريباً مشابه ساختار دستورات SQL است، اعمال مي­شود.

3- در اين مرحله بخش اصلي كار انجام مي­گيرد يعني بر اساس يك مدل مبتني بر MapReduce سعي مي­شود كه الگوهاي متناسب با محدوديت­هاي كاربران بر روي داده­هاي بزرگ استخراج شود و همچنين فضاي جست­وجو خيلي كوچكتر شود.

## فرضيه­ها و اهداف تحقيق

با توجه به موارد بررسي شده و چالش­هاي موجود در مورد زمينه تحقيق فرضيه­هايي مطرح شدند و متناسب با اين فرضيه­ها، اهداف اصلي تحقيق معرفي مي­شوند. در ادامه اين فرضيه­ها و اهداف معرفي مي­شوند.

* **فرضيه­ها:**

1- با ايجاد يك ساختار الگوي مكرر و اعمال علايق كاربران بر روي اين الگو، مي­توان فضاي جستجوي كاربران را در پايگاه داده­هاي عظيم به مراتب كوچكتر كرد.

2- با ايجاد يك مدل مبتني بر MapReduce در پايگاه داده­هاي عظيم مي­توان با دقت مناسبي الگوهاي مكرر دريافتي از كاربران را جستجو كرد.

* **اهداف:**

1- ايجاد يك فضاي جستجوي كوچك و متناسب با نياز و علايق كاربر در يك پايگاه داده عظيم

2- بيان علايق كاربر به صورت يك الگوي مشخص كه با دقت مناسبي بيان كننده نيازهاي كاربر باشد.

3- ايجاد يك روش كاوش مناسب الگوهاي كاربران داخل يك پايگاه داده عظيم مبتني بر مدل MapReduce

1. Frequent Pattern [↑](#footnote-ref-2)
2. Clustering [↑](#footnote-ref-3)
3. Classification [↑](#footnote-ref-4)
4. Anomaly Detection [↑](#footnote-ref-5)
5. Minimum Support [↑](#footnote-ref-6)
6. precise data [↑](#footnote-ref-7)
7. uncertain data [↑](#footnote-ref-8)
8. unvaried Uncertain data [↑](#footnote-ref-9)
9. Item Set Uncertain data [↑](#footnote-ref-10)
10. Tuple Uncertain data [↑](#footnote-ref-11)
11. Big Data [↑](#footnote-ref-12)
12. Big Data [↑](#footnote-ref-13)
13. Close Frequent Pattern [↑](#footnote-ref-14)
14. Maximal Frequent Pattern [↑](#footnote-ref-15)
15. Big Data [↑](#footnote-ref-16)
16. Big Data Mining [↑](#footnote-ref-17)
17. Seisint [↑](#footnote-ref-18)
18. LexisNexis [↑](#footnote-ref-19)
19. ChoicePoint [↑](#footnote-ref-20)
20. Hadoop Distributed File System [↑](#footnote-ref-21)
21. Minimum Support [↑](#footnote-ref-22)
22. -Maximal Frequent Pattern univariate Uncertain Data [↑](#footnote-ref-23)