42

Kilitlenme Tutarlılığı: FSCK ve Günlük Kaydı

Şimdiye kadar gördüğümüz gibi, dosya sistemi beklenen soyutlamaları uygulamak için bir takım veri yapısını yönetir: Dosyalar, dizinler ve bir dosya sisteminden beklediğimiz temel soyutlamayı desteklemek için gereken diğer tüm meta veriler. Çoğu veri yapısının aksine (örneğin, çalışan bir programın belleğinde bulunanlar), dosya sistemi veri yapıları **kalıcı(persist)** olmalıdır, yani, güç kaybına rağmen verileri tutan cihazlarda (sabit diskler veya flash tabanlı SSD’ler gibi) saklanarak uzun mesafe boyunca hayatta kalmaları gerekir.

Bir dosya sisteminin karşılaştığı en büyük zorluklardan biri, **güç kaybı (power loss)** veya **sistem kilitlenmesi (system crash)** olmasına ragmen kalıcı veri yapılarının nasıl güncelleneceğidir. Özellikle, disk yapılarını güncellemenin tam ortasında, birisi güç kablosuna takılırsa ve makine güç kaybederse? Veya işletim sistemi bir hatayla karşılaşır ve çökerse ne olur? Güç kayıpları ve kitlenmeler nedeniyle, kalıcı bir veri yapısını güncellemek oldukça zor olabilir ve dosya sistemi uygulamasında **kilitlenme tutarlılığı sorunu (crash-consistency problem)** olarak bilinen yeni ve ilginç bir soruna yol açar.

Bu sorunun anlaşılması oldukça basit. Belirli bir işlemi tamamlamak için iki disk üzerinde yapıyı, A ve B’yi güncelleştirmeniz gerektiğini düşünün. Disk aynı anda yalnızca tek bir isteğe hizmet verdiğinden, bu isteklerden biri önce diske ulaşacaktır (A veya B). Bir yazma işlemi tamamlandıktan sonra sistem çökerse veya güç kaybederse, diskteki yapı **tutarsız (inconsistent)** bir durumda bırakılır. Ve böylece, bütün dosya sistemlerinin çözmesi gereken bir sorunumuz var:

CRUX: KİLİTLENMELERE RAĞMEN DİSK NASIL GÜNCELLENİR

Sistem herhangi iki yazma işlemi arasında çökebilir veya güç kaybedebilir, ve bu nedenle diskteki durum yalnızca kısmen güncelleştirilebilir. Kilitlenmeden sonra, sistem önyüklenir ve dosya sistemini tekrar bağlamak ister (dosyalara ve benzerlerine erişmek için). Kilitlenmelerin zaman içinde rastgele noktalarda meydana gelebileceği göz önüne alındığında, dosya sisteminin diskteki görüntüyü makul bir durumda tutmasını nasıl sağlayabiliriz?

1

Bu bölümde, bu sorunu daha ayrıntılı olarak betimleyeceğiz ve dosya sistemlerinin üstesinden gelmek için kullandığı bazı yöntemlere bakacağız. **fsck** ya da **dosya sistemi kontrolcüsü (file system checker)** olarak bilinen eski dosya sistemleri tarafından benimsenen yaklaşımı inceleyerek başlayacağız. Daha sonra dikkatimizi **günlük kaydı (journaling)** (**önceden yazma günlüğü (write-ahead logging)** olarak da bilinir) olarak bilinen, her yazmaya biraz ek yük ekleyen ancak kilitlenmelerden veya güç kayıplarından daha hızlı kurtulan bir teknik olan başka bir yaklaşıma çevireceğiz. Linux ext3’ün [T98,PAA05] (nispeten modern bir günlük kaydı dosya sistemi) uyguladığı birkaç farklı günlük kaydı türü de dahil olmak üzere günlük kaydının temel mekanizmasını tartışacağız.

## Detaylı Bir Örnek

Günlük tutma araştırmamıza başlamak için bir örneğe bakalım. Disk üzeri yapıları bir şekilde güncelleyen bir **iş yükü (workload)** kullanmamız gerekir. Burada iş yükünün basit olduğunu farz edelim: Tek bir veri bloğunun mevcut bir dosyaya eklenmesi. Ekleme, dosya açarak ve dosya ofsetini dosyanın sonuna taşımak için lseek() öğesini çağırarak ve ardından kapatmadan önce dosyaya tek bir 4 KB yazma işlemi göndererek gerçekleştirilir.

Ayrıca, daha önce gördüğümüz dosya sistemlerine benzer şekilde, diskte standart basit dosya sistemi yapıları kullandığımızı farz edelim. Bu küçük örnek, bir **düğüm bit eşlemi (inode bitmap)** (her inode için bir tane olmak üzere yalnızca 8 bit), bir **veri bit eşlemi (data bitmap)** (ayrıca 8 bit, veri bloğu başına bir), inode ‘lar (toplam 8, 0’dan 7’ye kadar numaralandırılmış ve dört bloğa yayılmış), ve veri blokları (toplam 8, 0’dan 7’ye kadar numaralandırılmış). İşte bu dosya sisteminin bir diyagramı:

Bit eşlemleri(Bitmaps)

Düğüm(Inode) Veri(Data)Düğümler(Inodes) Veri Blokları(Data Blocks)

Da

I[v1]

0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7

Eğer resimdeki yapılara bakarsanız, inode bitmap’inde işaretlenen tek bir inode’ın (inode numarası 2) ve veri bir eşleminde de işaretlenen tek bir tahsis edilmiş veri bloğunu(veri bloğu 4)tahsis edildiğini görebiliriz. Bu inode‘ın ilk versiyonu olduğu için inode l[v1] olarak gösterilir; yakında güncellenecektir (yukarıda açıklanan iş yükü nedeniyle).

Bu basitleştirilmiş inode’un içine de göz atalım. l[v1]’in içinde şunları görüyoruz:

owner : remzi permissions : read-write size : 1

pointer : 4

pointer : null

pointer : null

pointer : null

Bu basitleştirilmiş inode’da, dosyanın boyutu 1’dir (bir bloğa tahsis edilmiştir), ve blok 4’e ilk doğrudan işaret işaret eder(dosyanın ilk veri bloğu, Da), ve diğer üç doğrudan işaretçinin tümü sıfır (kullanılmadıklarını gösterir) olarak ayarlanır. Elbette, gerçek inode’lerin daha çok alanları vardır; daha fazla bilgi için önceki bölümlere bakın.

Dosyaya eklediğimizde, dosyaya yeni bir veri bloğu ekliyoruz, ve böylece üç disk üstü yapıyı güncellememiz gerekiyor: İnode (yeni bloğa işaret etmeli ve ekleme nedeniyle yeni daha büyük boyutu kaydedilmelidir), yeni veri bloğu Db ve yeni veri bloğunun tahsis edildiğini belirtmek için veri bitmap’inin (bunu B[v2] olarak çağırın) yeni bir sürümü.

Böylece, sistemin hafızasında, diske yazmamız gereken üç bloğumuz var. Güncellenmiş İnode (inode versiyon 2 veya kısaca l[v2]) şimdi böyle gözüküyor:

owner : remzi permissions : read-write size : 2

pointer : 4

pointer : 5

pointer : null

pointer : null

Güncellenmiş veri bitmap’i(B[v2]) şimdi bunun gibi gözükür: 00001100. Son olarak, kullanıcıların dosyalara koyduğu herhangi bir şeyle dolu olan bir veri bloğu vardır. Belki de çalıntı müzik?

İstediğimiz dosya sisteminin diskteki son görüntüsünün böyle gözükmesi:

Bit eşlemleri(Bitmaps)

Düğüm(Inode) Veri(Data)Düğümler(Inodes) Veri Blokları(Data Blocks

I[v2]

0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7

Da Db

Bu geçişi başarmak için, dosya sistemi her bir inode (l[v2]), bitmap’i (B[v2]), ve veri bloğu (Db) diske üç parçada yazma işlemi gerçekleştirmelidir. Bu yazmaların genellikle kullanıcı write() sistem çağrısı yaptığında hemen gerçekleşmediğini unutmayın; bunun yerine, kirli inode, bitmap, ve yeni veriler önce bir süre ana bellekte (**sayfa önbelleğinde (page cache)** ya da **arabellek önbelleğinde (buffer cache))** durur; Daha sonra, dosya sistemi nihayet bunları diske yazmaya karar verdiğinde (5 saniye veya 30 saniye sonra), dosya sistemi gerekli yazma isteklerini diske gönderir. Ne yazık ki, kilitlenme meydana gelebilir ve bu nedenle diskteki bu güncelleştirmeleri engelleyebilir. Özellikle, bu yazmalardan bir veya ikisi gerçekleştikten sonra bir kilitlenme gerçekleşirse, ancak üçü birden gerçekleşmezse, dosya sistemi komik bir durumda bırakılabilir.

### Kilitlenme Senaryoları(Crash Scenarios)

Sorunu daha iyi anlamak için, bazı örnek kilitlenme senaryolarına bakalım. Sadece tek bir yazmanın başarılı olduğunu hayal edin; burada listelediğimiz üç olası ihtimal vardır:

* **Sadece veri bloğunun (Db) diske yazılması.** Bu durumda, veri disktedir, fakat bunu ona işaret eden bir inode yoktur ve blok tahsis edilmiş dese bile bitmap’i yoktur. Bu durumda, sanki hiç yazılmamış gibi olur. Bu durum dosya sistemi kilitlenme tutarsızlığı1 açısından hiç de sorun değildir.
* **Sadece güncellenmiş inode’un (l[v2]) diske yazılması.** Bu durumda, inode, Db’nin yazılmak üzere olduğu disk adresine (5) işaret eder, ancak Db henüz oraya yazılmamıştır. Bu nedenle, bu işaretçiye güvenirsek, **çöp (garbage)** verilerini diskten okuyacağız (disk adresi 5in eski içeriği).

Dahası, **dosya sistemi tutarsızlığı(file-system inconsistency)** dediğimiz, yeni bir sorunumuz var. Diskteki bitmap’i bize veri bloğu 5’in henüz tahsis edilmediğini söylüyor, fakat inode edildiğini söyler. Bitmap’i ve inode arasındaki uyuşmazlık bir dosya sisteminin veri yapılarındaki bir tutarsızlıktır; dosya sistemini kullanmak için bu sorunu bir şekilde çözmeliyiz(bununla ilgili daha fazla bilgi aşağıda)

* **Sadece güncellenmiş bitmap’inin (B[v2]) diske yazılması.** Bu durumda, bitmap’i blok 5’in ayrıldığını gösterir, ancak buna işaret eden bir inode yoktur. Böylece dosya sistemi yine tutarsızdır; çözülmeden bırakılırsa, blok 5 dosya sistemi tarafından asla kullanılmayacağından bu yazma bir **boşluk sızıntısına (space leak)** neden olur.

Diske üç blok yazma girişiminde üç kilitlenme senaryosu daha vardır. Bu gibi durumlarda, iki yazma başarılı olur ve sonuncusu başarısız olur:

* **İnode (l[v2]) ve bitmap’i (B[v2]) diske yazılır, ancak veri (Db) yazılmaz.** Bu durumda, dosya sistemi meta verileri tamamen tutarlıdır: İnode 5’i engellemek için bir işaretçiye sahiptir, bitmap’i 5’in kullanımda olduğunu gösterir ve böylece dosya sisteminin meta verilerinin açısından her şey yolunda gözükür. Ama bir sorun var: 5’in içinde yine çöp var.
* **İnode (l[v2]) ve veri bloğu (Db) yazılır, ancak bitmap’i (B[v2]) yazılmaz.** Bu durumda diskteki doğru verileri işaret eden inode’umuz var, ancak yine de inode ile bitmap’inin eski sürümü (B1) arasında bir tutarsızlık var. Bu nedenle, dosya sistemini kullanmadan önce sorunu bir kez daha çözmemiz gerekiyor.
* **Bitmap’i (B[v2]) ve veri bloğu (Db) yazılır, ancak inode (l[V2]) yazılmaz.** Bu durumda, yine inode ve veri bitmap’i arasında bir tutarsızlık var. Ancak, blok yazılmış olmasına rağmen ve bitmap’i kullanımını göstermesine rağmen, hiçbir inode dosyaya işaret etmediğinden, hangi dosyaya ait olduğu hakkında hiçbir fikrimiz yoktur.

1Ancak, bazı verileri kaybeden kullanıcı için bir sorun olabilir!

### Kilitlenme Tutarlılığı Sorunu(The Crash Consistency Problem)

Umarım, bu kilitlenme senaryolarından, çökme nedeniyle disk üzerindeki dosya sistemi görüntümüzde oluşabilecek birçok sorunu görebilirsiniz: Dosya sistemi veri yapılarında tutarsızlık olabilir; Boşluk sızıntılarımız olabilir; çöp verilerini bir kullanıcıya döndürebiliriz; ve benzeri. İdeal olarak yapmak istediğimiz şey, dosya sistemini tutarlı bir durumdan(örneğin, dosya eklenmeden önce) diğerine **atomik olarak (atomically)** (Örneğin, inode, bitmap’i ve yeni veri bloğu diske yazıldıktan sonra)olarak taşımaktır. Ne yazık ki, bunu kolayca yapamayız çünkü disk her seferde yalnızca bir yazma işlemi gerçekleştirir ve bu güncellemeler arasında kilitlenmeler ve güç kaybı oluşabilir. Bu genel soruna **kilitlenme-tutarlılık sorunu (crash-consistency problem)** diyoruz(buna **tutarlı-güncelleme sorunu (consistent-update problem)** da diyebiliriz).

## Çözüm #1: Dosya Sistemi Kontrolcüsü (The File System Checker)

İlk dosya sistemleri, kilitlenme tutarlılığına basit bir yaklaşım benimsedi. Temel olarak, tutarsızlıkların olmasına izin vermeye ve daha sonra bunları düzeltmeye karar verdiler. Bu tembel yaklaşımın klasik bir örneği, bunu yapan bir araçta bulunur: **fsck2**. Fsck, bu tür tutarsızlıkları bulmak ve onarmak için kullanılan bir UNIX aracıdır [MK96]; bir disk bölümünü kontrol etmek ve onarmak için benzer araçlar farklı sistemlerde mevcuttur. Böyle bir yaklaşımın tüm sorunları çözemeyeceğini unutmayın; örneğin, dosya sisteminin tutarlı göründüğü ancak inode’un çöp verilerine işaret ettiği yukarıdaki durumu düşünün. Tek gerçek amaç, dosya sistemi meta verilerinin dahili olarak tutarlı olduğundan emin olmaktır. fsck aracı, McKusick ve Kowalski’nin makalesinde özetlendiği gibi birkaç aşamada çalışır [MK96]. Dosya sistemi bağlanmadan ve kullanılabilir hale getirilmeden *önce* çalıştırılır (fsck, çalışırken başka hiçbir dosya sistemi etkinliğinin devam etmediğini varsayar); tamamlandığında, diskteki dosya sistemi tutarlı olmalı ve böylece kullanıcılar tarafından erişilebilir hale getirilebilir.

İşte Fsck’nın ne yaptığının temel bir özeti:

* **Süper Blok (SuperBlock):** fsck ilk önce süper bloğun makul görünüp görünmediğini kontrol eder, çoğunlukla dosya sistemi boyutunun tahsis edilen blok sayısından daha büyük olduğundan emin olmak gibi mantık kontrolleri yapar. Genellikle bu mantık kontrollerinin amacı, şüpheli (bozuk) bir süper blok bulmaktır; Bu durumda, sistem (veya yönetici) süper bloğun alternatif bir kopyasını kullanmaya karar verebilir.
* **Serbest Bloklar (Free blocks):** Daha sonra fsck, dosya sisteminde hangi blokların şu anda tahsis edildiğini anlamak için inode’lari, dolaylı blokları, çift dolaylı blokları vb. tarar. Bu bilgiyi, tahsis bitmap’lerinin doğru bir versiyonunu üretmek için kullanır; Böylece, bitmap’leri ve inode’lar arasında herhangi bir tutarsızlık varsa, inodelar içindeki bilgilere güvenerek çözülür. Tüm inodelar için aynı tür bir kontrol gerçekleşrilir ve kullanımda

2“eff-ess-see-kay”, “eff-ess-check” veya aracı beğenmezseniz “eff- suck” olarak telaffuz edilir. Evet, ciddi profesyonel insanlar bu terimi kullanır.

gibi görünen tüm inodelarin inode bitmap’lerinde bu şekilde işaretlendiğinden emin olunur.

* **Düğüm Durumu (Inode state):** Her inode bozulma veya başka sorunlar için kontrol edilir. Örneğin fsck, tahsis edilen her inode’un geçerli bir tür alanına sahip olduğundan emin olur (örneğin normal dosya, dizin, sembolik bağlantı, vb.). İnode alanlarıyla ilgili kolayca düzeltilemeyen sorunlar varsa, inode şüpheli olarak kabul edilir ve fsck tarafından temizlenir; inode bitmap’i buna uygun olarak güncellenir.
* **Düğüm Bağlantıları (Inode links):** fsck ayrıca tahsis edilen her inode ‘un bağlantı sayısını doğrular. Hatırlayabileceğiniz gibi, bağlantı sayısı bu belirli dosyaya bir referans (yani bir bağlantı) içeren farklı dizinlerin sayısını gösterir. Bağlantı sayısını doğrulamak için fsck, kök dizinden başlayarak tüm dizin ağacını tarar ve dosya sistemindeki her dosya ve dizin için kendi bağlantı sayılarını oluşturur. Yeni hesaplanan sayım ile bir inode içinde bulunan sayı arasında bir uyumsuzluk varsa, genellikle inode içindeki sayıyı sabitleyerek düzeltici önlem alınmalıdır. Tahsis edilmiş bir inode bulunursa, ancak hiçbir dizin ona atıfta bulunmazsa, *kayıp+bulunan(lost+found)* dizinine taşınır.
* **Yenilenenler (Duplicates):** fsck ayrıca yinelenen işaretçileri, yani iki farklı inode‘un aynı bloğa başvurduğu durumları da kontrol eder. Bir inode açıkça kötüyse, silinebilir. Alternatif olarak, işaret edilen blok kopyalanabilir, böylece her inode istenildiği gibi kendi kopyasını verir.
* **Kötü Bloklar (Bad Blocks):** Tüm işaretçilerin listesi taranırken hatalı blok işaretçileri için bir kontrol de gerçekleştirilir. Bir işaretçi geçerli aralığının dışındaki bir şeye açıkça işaret ediyorsa, örneğin bölüm boyutundan daha büyük bir bloğa atıfta bulunan bir adrese sahipse “kötü” olarak kabul edilir. Bu durumda, fsck çok zekice bir şey yapamaz; sadece işaretçiyi inode veya dolaylı yoldan kaldırır (temizler).
* **Dizin Kontrolleri (Directory checks):** fsck kullanıcı dosyalarının içeriğini anlamaz; ancak, dizinler dosya sisteminin kendisi tarafından oluşturulan özel olarak biçimlendirilmiş bilgileri tutar. Böylece, fsck her dizinin içeriği üzerinde ek bütünlük kontrolleri yaparak, “.” ve “..” girişlerinin ilk girişler olduğundan, bir dizin girişinde başvurulan her inode‘un tahsis edildiğinden ve hiçbir dizinin tüm hiyerarşide birden fazla kez bağlanmadığından emin olur.

Gördüğünüz gibi, çalışan bir fsck oluşturmak, dosya sistemi hakkında karmaşık bilgi gerektirir; böyle bir kod parçasının her durumda çalıştığından emin olmak zor olabilir [G+08]. Ancak, fsck’nın (ve benzer yaklaşımların) daha büyük ve belki de daha temel sorunu vardır: onlar *çok yavaşlar*. Çok büyük bir disk birimiyle, ayrılan tüm blokları bulmak ve tüm dizin ağacını okumak için tüm diski taramak birkaç dakika veya saat sürebilir. Disklerin kapasitesi arttıkça ve RAID’lerin popülaritesi arttıkça fsck’nin performansı engelleyici hale geldi(son gelişmelere rağmen [M+13]). Daha yüksek bir seviyede, fsck’nın temel öncülü biraz mantıksız görünüyor. Diske yalnızca üç bloğun yazıldığını yukarıdaki örneğimizi düşünün; sadece üç blokluk bir güncelleme sırasında ortaya çıkan sorunları çözmek için tüm diski taramak inanılmaz derecede pahalıdır. Bu durum yatak odanızda anahtarlarınızı yere ve ardından bodrumdan başlayarak her odaya doğru ilerleyerek tüm-evi-anahtarlar-için-ara kurtarma algoritmasını başlatmaya benzer. İşe yarıyor fakat israf. Böylece, diskler (ve RAID’ler) büyüdükçe, araştırmacılar ve uygulayıcılar başka çözümler aramaya başladılar.

## Çözüm #2: Günlük Kaydı (veya İleriye Yazma Kaydı) ( Journaling ( or Write – Ahead Logging ))

Muhtemelen tutarlı güncelleme sorunun en popüler çözümü, veritabanı yönetim sistemleri dünyasından bir fikir çalmaktır. **Önceden yazma günlüğü(write-ahead logging)** olarak bilinen bu fikir, tam olarak bu tür bir sorunu çözmek için icat edildi. Dosya sistemlerinde, genellikle tarihsel nedenlerden dolayı önceden yazma **günlüğü(journaling)** olarak adlandırırız. Bunu yapan ilk dosya sistemi Cedar [H87] idi, ancak Linux ext3 ve ext4, reiserfs, SGI’nin XFS’si ve Windowsun NTFS’si dahil olmak üzere birçok modern dosya sistemi bu fikri kullanıyor.

Temel fikir aşağıdaki gibidir. Diski güncellerken, yapıların üzerine yazmadan önce, ne yapmak üzere olduğunuzu açıklayan küçük bir not yazın (diskte başka bir yerde, iyi bilinen bir yerde ). Bu notu yazmak “önceden yaz” kısmıdır ve biz de bunu “günlük (log)” olarak düzenle-diğimiz bir yapıya yazarız; bu nedenle, önceden yazma günlüğü.

Notu diske yazarak, güncellediğiniz yapıların güncelleştirilmesi (üzerine yazma) sırasında bir kilitlenme meydada gelirse, geri dönüp yaptığınız nota bakıp tekrar deneyebileceğinizi garanti etmiş oluyorsunuz; böylece, tüm diski taramak yerine, bir kilitlenmeden sonra tam olarak neyi düzelteceğinizi (ve nasıl düzelteceğinizi) bileceksiniz. Tasarım gereği, günlük kaydı bu nedenle kurtarma sırasında gereken iş miktarını büyük ölçüde azaltmak için güncellemeler sırasında biraz iş ekler.

Şimdi, popüler bir günlük kaydı dosya sistemi olan **Linux ext3’**ün günlük kaydını dosya sistemine nasıl dahil ettiğini açıklayacağız. Disk üstü yapıların çoğu **Linux ext2** ile aynıdır, Örneğin, disk blok gruplarına ayrılmıştır ve her blok grubu bir inode bir eşleniği, veri bir eşleniği, inode’lar ve veri blokları içerir. Yeni anahtar yapısı, bölüm içinde veya başka bir cihazda az miktarda yer kaplayan günlüğün kendisidir. Böylece, bir ext2 dosya sistemi (günlük kaydı olmadan) şöyle görünür:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Süper | Grup 0 | Grup 1 | . . . | Grup N |  |

Günlüğün aynı dosya sistemi görüntüsüne yerleştirildiğini varsayarsak (ancak bazen ayrı bir aygıtta veya dosya sistemi içinde bir dosya olarak yerleştirilir), günlüğü olan bir ext3 dosya sistemi şöyle görünür:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Süper | Günlük | Grup 0 | Grup 1 | . . . | Grup N |  |

Gerçek fark, günlüğün varlığı ve elbette nasıl kullanıldığıdır.

### Veri Günlüğü ( Data Journaling )

**Veri günlüğü’nün (data journaling)** nasıl çalıştığını anlamak için basit bir örneğe bakalım. Veri günlüğü, bu tartışmanın çoğunun dayandığı Linux ext3 dosya sistemi ile bir modifikasyon(mod) olarak kullanılabilir.

Diyelim ki inode (l[v2]), bitmap’i (B[v2]) ve veri bloğunu (Db) tekrar diske yazmak istediğimiz kanonik bir güncellememizi tekrar aldık. Onları son disk konumlarına yazmadan önce, şimdi onları log’a (diğer adıyla günlük) yazacağız. Bu günlükte şöyle görünecektir:

Günlük

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TxB | I[v2] | B[v2] | Db | TxE |  |

Burada beş blok yazdığımızı görebilirsiniz. İşlem başlangıcı (transaction begin (TxB)), dosya sisteminde bekleyen güncelleme (Örneğin, l[v2], B[v2] ve Db bloklarının son adresleri) hakkında bilgiler ve bir tür **İşlem tanımlayıcı (IT) (transaction identifier (TID))** da dahil olmak üzere bize bu güncelleştirme hakkında bilgi verir. Orta üç blok sadece blokların kendilerinin tam içeriğini içerir; bu, güncelleştirmenin tam fiziksel içeriğini günlüğe koyduğumuz için **fiziksel günlük kaydı (physical logging)** olarak bilinir (alternatif bir fikir, **mantıksal günlük kaydı(logical logging)**, güncelleştirmenin daha kompakt bir mantıksal temsilini günlükte gösterir, Örneğin, “bu güncelleme, X dosyasına Db’yi eklemek istiyor”, bu biraz daha karmaşıktır, ancak günlükte yer tasarrufu sağlayabilir ve belki de performansı artırabilir). Son blok (final block (TxE)), bu işlemin sonunun bir göstergesidir ve ayrıca IT’yi de içerecektir.

Bu işlem güvenli bir şekilde diskte olduğunda, dosya sistemindeki eski yapıların üzerine yazmaya hazırız; bu işleme **kontrol noktası (checkpointing)** denir. Bu nedenle, dosya sistemini **kontrol (checkpoint)** etmek için (yani, günlükte bekleyen güncelleme ile güncel hale getirin), yukarıda görüldüğü gibi I[v2], B[v2], ve Db yazmalarını disk konumuna veriyoruz; bu yazmalar başarıyla tamamlanırsa, dosya sistemini başarıyla kontrol etmiş oluruz ve temel olarak işimiz biter. Böylece, ilk işlem sıramız:

1. **Günlük Yazma (Journal write):** İşlem başlangıç bloğu, bekleyen tüm veriler ve meta veri güncellemeleri ve bir işlem bitiş bloğu dahil olmak üzere işlemi günlüğe yazma; Bu yazmaların tamamlanmasını bekleyin.
2. **Kontrol Noktası (Checkpoint):** Bekleyen meta verileri ve veri güncellemelerini dosya sistemindeki son konumlarına yazma.

Örneğimizde, günlüğe önce TxB, l[v2], B[v2], Db ve TxE yazacağız. Bu yazma işlemleri tamamlandığında, I[v2], B[v2], ve Db’yi güncelleştirmeyi tamamlarız.

Günlüğe yazma sırasında bir kilitlenme meydana gelirse işler biraz daha zorlaşır. Burada, işlemdeki blok kümesini (örneğin, TxB, I[v2], B[v2], Db, TxE) diske yazmaya çalışıyoruz. Bunu yapmanın basit bir yolu, her birini bir kerede yayınlamak, her birinin tamamlanmasını beklemek ve ardından bir sonrakini yayınlamak olacaktır. Ancak, bu yavaştır.

BIR YANA: DISKE YAZMAYI ZORLAMA

İki disk yazma arasında sıralamayı zorlamak için, modern dosya sistemlerinin birkaç ekstra önlem alması gerekir. Eski zamanlarda, iki yazma arasında sıralamayı zorlamak, A ve B, kolaydı: sadece A’nın diske yazılmasını sağlayın, yazma tamamlandığında diskin işletim sistemini kesmesini bekleyin ve ardından B’nin yazmasını sağlayın.

Disklerdeki yazma önbelleklerinin artan kullanımı nedeniyle işler biraz daha karmaşıklaştı. Yazma arabelleğe alma etkinken (bazen **anında raporlama (immediate reporting)** olarak adlandırılır), bir disk, diskin bellek önbelleğine yerleştirildiğinde ve henüz diske ulaşmadığında, işletim sistemine yazma işleminin tamamlandığını bildirir. İşletim sistemi daha sonra bir sonraki yazma işlemini yayınlarsa, önceki yazmalardan sonra diske ulaşacağı garanti edilmez; böylece yazılar arasındaki sıralama korunmaz. Çözümlerden biri, yazma arabelleğe almayı devre dışı bırakmaktır. Fakat, daha modern sistemler ekstra önlemler alır ve açık **yazma engelleri (write barriers)** çıkarır; Böyle bir engel, tamamlandığında, engelden önce verilen tüm yazmaların, engelden sonra verilen herhangi bir yazmadan önce diske ulaşmasını garanti eder.

Tüm bu makineler, diskin doğru çalışması için büyük bir güven gerektirir. Ne yazık ki, son araştırmalar bazı disk üreticilerinin, “daha yüksek performanslı” diskler sunmak amacıyla, yazma engeli isteklerini açıkça görmezden geldiğini ve böylece disklerin daha hızlı çalışmasını sağladığını ancak yanlış çalışma riski altında olduğunu gösteriyor [C+13, R+11]. Kahan’ın dediği gibi, hızlı yanlış olsa bile, neredeyse her zaman yavaş olanı yener.

İdeal olarak, beş blok yazmanın tümünü bir kerede yayınlamak isteriz, çünkü bu, beş yazmayı tek bir sıralı yazıya dönüştürür ve böylece daha hızlı olur. Ancak, aşağıdaki nedenlerden dolayı bu güvenli değildir: Bu kadar büyük bir yazma göz önüne alındığında, disk dahili olarak zamanlama yapabilir ve büyük yazmanın küçük parçalarını herhangi bir sırayla tamamlayabilir. Böylece, disk dahili olarak (1) TxB, I[v2], B[v2], ve TxE yazabilir ve ancak daha sonra Db yazabilir. Ne yazık ki, disk (1) ve (2) arasında güç kaybederse, diskte sona eren şey budur:

Günlük

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TxB  id=1 | I[v2] | B[v2] | ?? | TxE  id=1 |  |

Bu neden bir problem? Peki, işlem geçerli bir işlem gibi görünüyor (eşleşen sıra numaralarıyla bir başlangıcı ve sonu var). Dahası, dosya sistemi bu dördüncü bloğa bakamaz ve yanlış olduğunu bilemez; sonuçta, keyfi kullanıcı verileridir. Böylece, sistem şimdi yeniden başlatılır ve kurtarma işlemini çalıştırırsa, bu işlemi yeniden yürütecek ve ‘??’ çöp bloğunun içeriğini Db’nin yaşaması gereken konuma cahilce kopyalar. Bu, bir dosyadaki keyfi kullanıcı verileri için kötüdür; dosya sistemini monte edilemez hale getirebilecek süper blok gibi kritik bir dosya sistemi parçasına olursa çok daha kötü olur

BİR YANA: GÜNLÜK YAZMALARINI OPTİMİZE ETMEK

Günlüğe yazmanın belirli bir verimsizliğini fark etmiş olabilirsiniz. Yani, dosya sistemi önce işlem-başlangıç bloğunu ve işlemin içeriğini yazmalıdır; ancak bu yazma işlemleri tamamlandıktan sonra dosya sistemi işlem-sonu bloğunu diske gönderebilir. Bir diskin nasıl çalıştığını düşünürseniz, performans etkisi açıktır: Genellikle fazladan bir rotasyon meydana gelir (nedenini düşünün).

Eski yüksek lisans öğrencilerimizden biri olan Vijayan Prabhakaran’ın, bu sorunu çözmek için basit bir fikri vardı [P+05]. Günlüğe bir işlem yazarken, günlüğün içeriğinin bir sağlama toplamını başlangıç ve bitiş bloklarına ekleyin. Bunu yapmak, dosya sisteminin beklemeye gerek kalmadan tüm işlemi bir kerede yazmasını sağlar; kurtarma sırasında, dosya sistemi hesaplanan sağlama toplamında işlemde depolanan sağlama toplamı ile uyumsuzluk görürse, işlemin yazılması sırasında bir kilitlenme oluştuğu sonucuna varabilir ve böylece dosya-sistemi güncellemesi atabilir. Böylece, yazma protokolünde ve kurtarma sisteminde küçük bir ince ayar ile, dosya sistemi daha hızlı ortak-durum performansı elde edebilir; bunun da ötesinde, günlükten yapılan tüm okumalar artık bir sağlama toplamı tarafından korunduğundan sistem biraz daha güvenilirdir.

Bu basit düzeltme Linux dosya sistemi geliştiricilerinin dikkatini çekecek kadar çekiciydi ve daha sonra onu (tahmin ettiniz!) **Linux ext4** adlı yeni nesil Linux dosya sistemine dahil ettiler. Artık Android el tipi platform da dahil olmak üzere dünya çapında milyonlarca makinede gönderim yapıyor. Böylece, birçok Linux tabanlı sistemde diske her yazdığınızda, Wisconsin’de geliştirilen küçük bir kod, sisteminizi biraz daha hızlı ve güvenilir hale getirir

Bu sorunu önlemek için, dosya sistemi işlemsel yazmayı iki adımda gerçekleştirir. İlk olarak, TxE bloğu dışındaki tüm blokları günlüğe yazar, bu yazıların hepsini bir kerede yayınlar. Bu yazmalar tamamlandığında, günlük şuna benzer (ekleme iş yükümüzü yeniden varsayarsak):

Günlük

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TxB  id=1 | I[v2] | B[v2] | Db |  |

Bu yazmalar tamamlandığında, dosya sistemi TxE bloğunun yazımını yayınlar ve böylece günlüğü bu son, güvenli durumda bırakır:

Günlük

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TxB  id=1 | I[v2] | B[v2] | Db | TxE  id=1 |  |

Bu işlemin önemli bir yönü, disk tarafından sağlanan atomiklik garantisidir. Diskin, herhangi bir 512 baytlık yazmanın gerçekleşeceğini

veya gerçekleşmeyeceğini (ve asla yarı yazılmayacağını) garanti ettiği ortaya çıktı; bu nedenle, TxE’nin yazılmasının atomik olduğundan emin olmak için, onu tek bir 512 baytlık blok haline getirmelisiniz. Böylece, dosya sistemini güncellemek için mevcut protokolümüz, üç aşamasında her biri etiketlenmiş olarak:

* 1. **Günlük Yazma (Journal write):** İşlemin içeriğini (TxB, meta veriler, ve veriler dahil) günlüğe yazın; bu yazmaların tamamlanmasını bekleyin.
  2. **Günlük İşlenme (Journal commit):** İşlem taahhüt bloğunu (TxE içeren) günlüğe yazın; yazmanın tamamlanmasını bekleyin; işlemin **işlendiği (committed)** söylenir.
  3. **Kontrol Noktası (Checkpoint):** Güncellemenin içeriğini (meta veriler ve veriler) diskteki son konumlarına yazın.

### Kurtarma ( Recovery )

Şimdi bir dosya sisteminin bir kilitlenmeden **kurtulmak (recover)** için günlüğün içeriğini nasıl kullanabileceğini anlayalım. Bu güncelleştirme dizisi sırasında herhangi bir zamanda kilitlenme meydana gelebilir. Kilitlenme, işlem günlüğüne güvenli bir şekilde yazılmadan önce (yani, yukarıdaki Adım 2 tamamlanmadan önce) gerçekleşirse, o zaman işimiz kolaydır: bekleyen güncelleme basitçe atlanır. Kilitlenme, işlem günlüğe işlendikten sonra ancak kontrol noktası tamamlanmadan önce gerçekleşirse, dosya sistemi güncelleştirmeyi aşağıdaki gibi **kurtarabilir (recover)**. Sistem önyüklendiğinde, dosya sistemi kurtarma işlemi günlüğü tarar ve diske kaydedilmiş işlemleri tarar; Böylece bu işlemler **yeniden oynatılır (replayed)** (sırayla), dosya sistemi yeniden işlemdeki blokları diskteki son konumlarına yazmaya çalışır. Bu günlüğe kaydetme biçimi, var olan en basit biçimlerden biridir ve **yineleme günlüğü (redo logging)** olarak adlandırılır. Dosya sistemi, günlükte işlenen işlemleri kurtararak, diskteki yapıların tutarlı olmasını sağlar ve böylece dosya sistemini bağlayarak ve kendisini yeni istekler için hazırlayarak devam edebilir.

Kontrol noktası sırasında herhangi bir noktada, blokların son konumlarına yapılan bazı güncellemeler tamamlandıktan sonra bile bir kilitlenmenin meydana gelmesinin sorun olmadığını unutmayın. En kötü durumda, bu güncellemelerden bazıları kurtarma sırasında basitçe tekrar gerçekleştirilir. Kurtarma nadir görülen bir işlem olduğundan (yalnızca beklenmeyen bir sistem kilitlenmesinden sonra gerçekleşir), birkaç gereksiz yazma işlemi endişelenecek bir şey değildir3.

### Toplu Günlük Güncellemeleri ( Batching Log Updates)

Temel protokolün çok fazla ekstra disk trafiği ekleyebileceğini fark etmiş olabilirsiniz. Örneğin, aynı dizinde dosya1 ve dosya2 adlı bir satırda iki dosya oluşturduğumuzu düşünün.

3Her şey için endişelendiğiniz sürece, bu durumda size yardımcı olamayız. Bu kadar endişelenmeyi bırak, bu sağlıksız! Ama şimdi muhtemelen aşırı endişelenme konusunda endişeleniyorsun.

Bir dosya oluşturmak için, en az aşağıdakiler de dahil olmak üzere bir dizi disk yapınışı güncellemek gerekir: inode bitmap’i (yeni bir inode tahsis etmek için), dosyanın oluşturulan inode’u, yeni dizin girişini içeren üst dizinin veri bloğu ve (artık yeni bir değişiklik zamanına sahip) üst dizin inode’u. Günlük kaydı ile, tüm bu bilgileri mantıksal olarak iki dosya oluşturma işlemimizin her biri için günlüğe işleriz; çünkü dosyalar aynı dizindedir ve hatta aynı inode bloğu içinde inode’ları olduğunu varsayarsak, bu eğer dikkatli olmazsak, aynı blokları tekrar tekrar yazacağımız anlamına gelir.

Bu sorunu çözmek için, bazı dosya sistemleri her güncellemeyi diske teker teker işlemez (örneğin, Linux ext3); bunun yerine, tüm güncellemeleri küresel bir işlemde arabelleğe alabilirsiniz. Yukarıdaki örneğimizde, iki dosya oluşturulduğunda, dosya sistemi sadece bellek içi inode bir eşleniğini, dosyaların inode’larini, dizin verilerini ve dizin inode’unu kirli olarak işaretler, ve bunları geçerli işlemi oluşturan bloklar listesine ekler. Sonunda bu blokları diske yazma zamanı geldiğinde (örneğin, 5 saniyelik bir zaman aşımından sonra), yukarıda açıklanan tüm güncellemeleri içeren bu tek genel işlem gerçekleştirilir. Böylece, güncelleştirmeleri arabelleğe alarak, bir dosya sistemi çoğu durumda diske aşırı yazma trafiğini önleyebilir.

### Günlüğü Sonlu Hale Getirme ( Making The Log Finite )

Böylece, disk üzerindeki dosya sistemi yapılarını güncellemek için temel bir protokole ulaştık. Dosya sistemi, bellekteki güncellemeleri bir süre arabelleğe alır; Nihayet diske yazma zamanı geldiğinde, dosya sistemi önce işlemin ayrıntılarını dikkatlice günlüğe yazar (diğer adıyla ileri yazma günlüğü); işlem tamamlandıktan sonra, dosya sistemi bu blokları diskteki son konumlarına göre kontrol eder.

Ancak, günlük sonlu bir boyuttadır. Üzerine işlemleri eklemeye devam edersek (bu şekilde olduğu gibi), yakında dolacaktır. Sizce o zaman ne olur?

Günlük

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tx1 | Tx2 | Tx3 | Tx4 | Tx5 | ... |

Günlük dolduğunda iki sorun ortaya çıkar. İlki daha basit ama daha az kritik: Günlük ne kadar büyükse, kurtarma işleminin kurtarmak için günlükteki tüm işlemleri (sırayla) yeniden yürütmesi gerektiğinden, kurtarma işlemi o kadar uzun sürer. İkincisi daha çok sorun: Günlük dolu olduğunda (veya neredeyse doluyken), diske başka işlem yapılamaz, böylece dosya sistemi “kullanışlıdan daha az” (yani işe yaramaz) hale gelir.

Bu sorunları gidermek için, günlük kaydı dosya sistemleri günlüğü dairesel bir veri yapısı olarak ele alır ve günlüğü tekrar tekrar kullanır; bu nedenle günlük bazen **dairsel günlük (circular log)** olarak anılır. Bunu yapmak için, dosya sistemi bir kontrol noktasından bir süre sonra harekete geçmesi gerekir. Özellikle, bir işlem kontrol noktasından geçirildikten sonra, dosya sistemi, günlükte kapladığı alanı boşaltmalı ve günlük alanının

yeniden kullanılmasına izin vermelidir. Bu amaca ulaşmanın birçok yolu var; örneğin, günlükteki en eski ve en yeni kontrol noktası olmayan işlemleri bir **günlük süper bloğunda (journal superblock)** işaretleyebilirsiniz; diğer tüm alanlar boştur. İşte grafiksel bir tasvir:

Günlük



...

Tx5

Tx4

Tx3

Tx2

Tx1

Günlük Süper

Günlük süper bloğunda (ana dosya sistemi süper bloğu ile karıştırılmamalıdır), günlük kaydı sistemi, hangi işlemleri henüz kontrol edilmediğini bilmek için yeterli bilgiyi kaydeder ve böylece kurtarma süresini kısaltır ve günlüğün dairesel bir şekilde yeniden kullanılmasını sağlar. Ve böylece temel protokolümüze bir adım daha ekliyoruz:

1. **Günlük Yazma (Journal Write):** İşlemin içeriğini (TxB ve güncellemenin içeriğini içeren) günlüğe yazın; bu yazmaların tamamlanmasını bekleyin.
2. **Günlük İşlenme (Journal commit):** İşlem taahhüt bloğunu (TxE içeren) günlüğe yazın; yazmanın tamamlanmasını bekleyin; işlem artık **işlenmiştir (committed)**.
3. **Kontrol Noktası (Checkpoint):** Güncellemenin içeriğini dosya sistemindeki son konumlara yazın.
4. **Boş (Free):** Bir süre sonra, günlük süper bloğunu güncelleyerek işlemi günlükte boş olarak işaretleyin.

Böylece son veri günlüğü protokolümüze sahibiz. Fakat hala bir sorun var: her veri bloğunu diske *iki kez* yazıyoruz, bu da özellikle sistem kilitlenmesi gibi nadir bir şey için ödenmesi gereken ağır bir maliyettir. Verileri iki kez yazmadan tuturlılığı korumanın bir yolunu bulabilir misiniz?

### Meta Veri Günlüğe Kaydetme ( Metadata Journaling )

Kurtarma artık hızlı olmasına rağmen (tüm diski taramak yerine günlüğü tarama ve birkaç işlemi yeniden yürütme), dosya sisteminin normal çalışması arzu edebileceğimizden daha yavaştır. Özellikle, diske her yazma için, şimdi önce günlüğe yazıyoruz, böylece yazma trafiğini ikiye katlıyoruz; Bu ikiye katlama, artık sürücünün en yüksek yazma bant genişliğinin yarısında devam edecek olan sıralı yazma iş yükleri sırasında özellikle acı vericidir. Ayrıca, günlüğe yazmalar ile ana dosya sistemine yazmalar arasında, bazı iş yükleri gözle görülür ek yük oluşturan maliyetli bir arama vardır.

Her veri bloğunu diske iki kez yazmanın maliyeti nedeniyle, insanlar performansı hızlandırmak için birkaç farklı şey denediler. Örneğin, yukarıda tanımladığımız günlüğe kaydetme modu, tüm kullanıcı verilerini günlüğe kaydettiğinden (dosya sisteminin meta verilerine ek olarak) genellikle **veri günlük kaydı (data journaling)** olarak adlandırılır (Linux ext3’te olduğu gibi). Daha basit (ve daha yaygın) bir günlük tutma biçimine bazen **sıralı günlük kaydı(ordered journaling)** (veya sadece **meta veri**

**günlük kaydı(metadata journaling)**) denir ve kullanıcı verilerinin günlüğe *yazılmaması* dışında neredeyse aynıdır. Böylece, yukarıdaki güncellemenin aynısını yaparken, aşağıdaki bilgiler günlüğe yazılacaktır.

Günlük

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TxB | I[v2] | B[v2] | TxE |  |

Daha önce günlüğe yazılan veri bloğu Db, bunun yerine dosya sistemine uygun şekilde yazılır ve fazladan yazmadan kaçınır; diske gelen G/Ç trafiğinin çoğunun veri olduğu göz önüne alındığında, verileri iki kez yazmamak, günlük kaydının G/Ç yükünü önemli ölçüde azaltır. Yine de değişiklik ilginç bir soruyu gündeme getiriyor: Veri bloklarını diske ne zaman yazmalıyız?

Sorunu daha iyi anlamak için bir dosya ekleme örneğimizi tekrar ele alalım. Güncelleme üç bloktan oluşur: I[v2], B[v2], ve Db. İlk ikisi hem meta verilerdir hem de günlüğe kaydedilir ve ardından kontrol noktası işaretlenir; ikincisi dosya sistemine yalnızca bir kez yazılacaktır. Db’yi diske ne zaman yazmalıyız? Önemli mi?

Görünüşe göre, veri yazmanın sırası yalnızca meta veriler günlük kaydı için önemlidir. Örneğin, işlem tamamlandıktan *sonra* (I[v2] ve B[v2] içeren) diske Db yazarsak nolur? Ne yazık ki, bu yaklaşımın bir sorunu var: Dosya sistemi tutarlıdır, ancak I[v2] çöp verilerine işaret edebilir. Özellikle, I[v2] ve B[v2] yazıldığı ancak Db’nin diske girmediği durumu düşünün. Dosya sistemi daha sonra kurtarmaya çalışacaktır. Db günlükte *olmadığı* için, dosya sistemi yazmaları I[v2] ve B[v2]’ye yeniden yürütür ve tutarlı bir dosya sistemi üretir (dosya sistemi meta verileri açısından). Ancak, I[v2] çöp verilerine, yani, Db’nin yöneldiği yuvada ne varsa ona işaret ediyor olacak.

Bu durumun ortaya çıkmamasını sağlamak için, bazı dosya sistemleri (örneğin, Linux ext3) ilgili meta veriler diske yazılmadan *önce*, önce diske veri bloklarını (normal dosyalardan) yazar. Özellikle, protokol aşağıdaki gibidir:

* 1. **Veri Yazma (Data write):** Verileri son konuma yazın; tamamlanmasını bekleyin (bekleme isteğe bağlıdır; ayrıntılar için aşağıya bakın).
  2. **Günlük Meta Verisi Yazma (Journal metadata write):** Başlangıç bloğunu ve meta verileri günlüğe yazın; yazma işlemlerinin tamamlanmasını bekleyin.
  3. **Günlük İşlenme (Journal commit):** İşlem taahhüt bloğunu (TxE içeren) günlüğe yazın; yazmanın bitmesini bekleyin; işlem (veriler dahil) artık **işlenmiştir (committed)**.
  4. **Kontrol Noktası Meta Verileri (Checkpoint metadata):** Meta veri güncellemesinin içeriğini dosya sistemindeki son konumlara yazın.
  5. **Boş (Free):** Daha sonra, günlük süper bloğunda işlemi boş olarak işaretleyin.

Önce veri yazmayı zorlayarak, bir dosya sistemi bir işaretçinin asla çöpe işaret etmeyeceğini garanti edebilir. Gerçekten de, bu “işaret edilen nesneyi, ona işaret eden nesneden önce yazın” kuralı, kilitlenme tutarlılığının merkezinde yer alır ve diğer kilitlenme tutarlılığı şemaları [GP94] tarafından daha da sömürülür (ayrıntılar için aşağıya bakın).

Çoğu sistemde, meta veri günlük kaydı (ext3’ün sıralı günlük kaydına benzer) tam veri günlük kaydından daha popülerdir. Örneğin, Windows NTFS ve SGI’ın XFS bir tür meta veri günlük kaydı kullanır. Linux ext3 size veri, sıralı veya sırasız modlardan birini seçme seçeneği sunar (sırasız modda, veriler istediğiniz zaman yazılabilir). Tüm bu modlar, meta verileri tutarlı tutar; veri anlamlarında farklılık gösterirler.

Son olarak, yukarıdaki protokolde belirtildiği gibi, günlüğe yazma işlemi yapmadan (adım 2) önce veri yazmayı tamamlamaya zorlamanın (adım 1) doğruluk için gerekli olmadığını unutmayın. Özellikle, aynı anda verilere, işlem başlangıç bloğuna ve günlüğe kaydedilen meta verilere yazma işlemleri yayınlamak iyi olur; tek gerçek gereklilik, adım 1 ve 2’nin günlük taahhüt bloğunun yayınlanmasından önce tamamlanmasıdır (adım 3).

### Zor Durum: Yeniden Kullanımı Engelle ( Tricky Case : Block Reuse )

Günlük kaydını daha zor hale getiren ve bu nedenle tartışmaya değer bazı ilginç köşe durumları vardır. Bunların bir kısmı blok yeniden kullanımı etrafında döner; Stephen Tweedie’nin (ext3’ün arkasındaki ana güçlerden biri) dediği gibi:

“Tüm sistemin korkunç kısmı nedir? ... dosyaları siliyor. Silmekle ilgili her şey tehlikeli. Silme ile ilgili her şey... bloklar silinir ve sonra yeniden tahsis edilirse ne olacağına dair kabuslar görürsünüz” [T00]

Tweedie’nin verdiği özel örnek aşağıdaki gibidir. Bir tür meta veri günlüğü kullandığınızı varsayalım (ve bu nedenle dosyalar için veri blokları günlüğe *kaydedilmez*). Diyelim ki foo adlı bir dizininiz var. Kullanıcı foo’ya bir girdi ekler (diyelim ki bir dosya oluşturarak), ve böylece foo’nun içeriği (çünkü dizinler meta veri olarak kabul edilir) günlüğe yazılır; foo dizin verilerinin konumunun blok 1000 olduğunu varsayalım. Bu nedenle günlük, buna benzer bir şey içerir:

Günlük

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TxB  id=1 | I[foo]  ptr:1000 | D[foo]  [final addr:1000] | TxE  id=1 |  |

Bu noktada, kullanıcı dizindeki her şeyi ve kendisini siler ve 1000 numaralı bloğu yeniden kullanım için serbest bırakır. Son olarak, kullanıcı yeni bir dosya (bar diyelim) oluşturur ve bu da foo’ya ait olan aynı bloğu (1000) yeniden kullanılır. Bar inode’u, verileri gibi diske de bağlıdır; Bununla birlikte, meta veri günlük kaydı kullanımda olduğundan, günlüğe yalnızca bar’ın inode’u taahhüt edilir; bar dosyasındaki 1000 numaralı blokta yeni yazılan veriler günlüğe *kaydedilmez*.

Günlük

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TxB  id=1 | I[foo]  ptr:1000 | D[foo]  [final addr:1000] | TxE  id=1 | TxB  id=2 | I[bar]  ptr:1000 | TxE  id=2 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TxB | **Günlük**  İçerik  (meta veri) (veri) | TxE | **Dosya sistemi**  Meta veri Veri | |
| **İşlem tamamlandı** | **İşlem İşlem**  **tamamlandı**  **tamamlandı** |  |  | |
| **İşlem**  **tamamlandı** | | |  | |
|  | | | **İşlem**  **tamamlandı** | **İşlem**  **tamamlandı** |

Figure 42.1: **Veri günlüğü kaydı zaman çizelgesi**

Şimdi bir kilitlenme oluştuğunu ve tüm bilgilerin hala günlükte olduğunu varsayalım. Yeniden oynarma sırasında, kurtarma işlemi, blok 1000’ deki dizin verilerinin yazılması da dahil olmak üzere günlükteki her şeyi yeniden oynatır; böylece yeniden oynatma, mevcut bar dosyasının kullanıcı verilerinin üzerine eski dizin içeriklerini yazar! Açıkçası bu doğru bir kurtarma işlemi değildir ve kesinlikle bar dosyasını okurken kullanıcı için bir sürpriz olacaktır.

Bu sorunun bir dizi çözümü var. Örneğin, söz konusu blokların silinmesi günlükten kontrol noktasıyla işaretlenene kadar bloklar asla yeniden kullanılmaz. Linux ext3’ün bunun yerine yaptığı şey, günlüğe **iptal (revoke)** kaydı olarak bilinen yeni bir kayıt türü eklemektir. Yukarıdaki durumda, dizinin silinmesi günlüğe bir iptal kaydının yazılmasına neden olur. Günlüğü tekrar oynatırken, sistem önce bu tür iptal kayıtlarını tarar; Bu tür iptal edilen veriler asla tekrar oynatılmaz, böylece yukarıda belirtilen sorundan kaçınılır.

### Günlük Kaydını Tamamlama: Bir Zaman Çizelgesi

### (Wrapping Up Journaling: A Timeline)

Günlük kaydı tartışmamıza son vermeden önce, tartıştığımız protokollerin her birini gösteren zaman çizelgeleriyle özetliyoruz. Şekil 42.1 verileri ve meta verileri günlüğe kaydeden protokolü gösterirken, şekil 42.2 yalnızca meta verileri günlüğe kaydeden protokolü gösterir.

Her şekilde, zaman aşağı yönde artar ve şekilde her satır, bir yazma işleminin yapılabileceği veya tamamlanabileceği mantıksal zamanı gösterir. Örneğin, veri günlüğe kaydetme protokolünde (Şekil 42.1), işlemin yazmaları başlangıç bloğu (TxB) ve işlemin içeriği mantıksal olarak aynı anda yayınlanabilir ve böylece herhangi bir sırada tamamlanabilir; ancak, işlem bitiş bloğuna yazma (TxE), söz konusu önceki yazılar tamamlana kadar yayınlanmamalıdır. Benzer şekilde, kontrol noktası verilere yazar ve meta veri blokları, işlem sonu bloğu işlenene kadar başlamaz. Yatay kesikli çizgiler, yazma sırası gereksinimlerine nerede uyulması gerektiğini gösterir.

Meta veri günlüğe kaydetme protokolü için benzer bir zaman çizelgesi gösterilir. Veri yazma işleminin mantıksal olarak işleme yazma

|  |  |  |  |
| --- | --- | --- | --- |
| TxB | **Günlük**  İçerik  (meta veri) | TxE | **Dosya Sistemi**  Meta veri Veri |
| **İşlem**  **tamamlandı** | **İşlem**  **tamamlandı** |  | **İşlem tamamlandı** |
| **İşlem**  **tamamlandı** | | |  |
|  | | | **İşlem tamamlandı** |

Figure 42.2: **Meta veri Günlük Kaydı Zaman Çizelgesi**

işlemlerinin başlaması ve günlüğün içeriği ile aynı anda verilebileceğini unutmayın; ancak, işlem sona ermeden önce düzenlenmeli ve tamamlanmalıdır.

Son olarak, zaman çizelgelerindeki her yazma için işaretlenen tamamlanma süresinin keyfi olduğunu unutmayın. Gerçek bir sistemde, tamamlanma süresi, performansı artırmak için yazmaları yeniden sıralayabilen G/Ç alt sistemi tarafından belirlenir. Sahip olduğumuz sıralamayla ilgili tek garanti, protokol doğruluğu için uygulanması gerekenlerdir (ve şekillerde yatay kesikli çizgilerle gösterilir).

## Çözüm #3: Diğer Yaklaşımlar ( Other Approaches )

Şimdiye kadar dosya sistemi meta verilerini tutarlı tutmak için iki seçenek açıkladık: fsck’ye dayanan tembel bir yaklaşım ve günlük kaydı olarak bilinen daha aktif bir yaklaşım. Ancak, bunlar sadece iki yaklaşım değildir. Esnek güncellemeler (Soft Updates) [GP94] olarak bilinen böyle bir yaklaşım, Ganger and Patt tarafından tanıtıldı. Bu yaklaşım, diskteki yapıların hiçbir zaman tutarsız bir durumda bırakılmamasını sağlamak için dosya sistemine tüm yazmaları dikkatli bir şekilde sıralar. Örneğin, ona işaret eden inode’dan *önce* diske işaret eden bir veri bloğunu yazarak, inode’un asla çöpe işaret etmemesini sağlarız; dosya sisteminin tüm yapıları için benzer kurallar türetilebilir. Bununla birlikte, esnek güncellemeri uygulamak zor olabilir; yukarıda açıklanan günlük kaydı katmanı, tam dosya sistemi yapıları hakkında nispeten az bilgi ile uygulanabilirken, esnek güncellemeler, her dosya sistemi veri yapısı hakkında karmaşık bilgi gerektirir ve bu nedenle sisteme makul miktarda karmaşıklık ekler.

Diğer bir yaklaşım **yazma üzerine kopyalama (copy-on-write)** (evet , **COW**), olarak bilinir ve Sun’ın ZFS [B07] de dahil olmak üzere bir dizi popüler dosya sisteminde kullanılır. Bu teknik hiçbir zaman dosyaların veya dizilerin üzerine yazmaz; bunun yerine, diskte daha önce kullanılmayan konumlara yeni güncellemeler yerleştirir. Bir dizi güncelleme tamamladıktan sonra, COW dosya sistemleri yeni güncellenen yapılara işaretçiler eklemek için dosya sisteminin kok yapısını çevirir. Bunu yapmak, dosya sisteminin tutarlı kalmasını kolaylaştırır. Gelecekteki bir bölümde günlük yapılandırılmış dosya sistemi

(log-structured file system) (LFS) tartışırken bu teknik hakkında daha fazla bilgi edineceğiz; LFS, bir COW’un erken bir örneğidir.

Başka bir yaklaşım da Wisconsin’de yeni geliştirdiğimiz bir yaklaşımdır. **Geri işaretçi tabanlı tutarlılık (backpointer-based consistency)** (veya **GİTT** (**BBC)**) başlıklı bu teknikte, yazmalar arasında sıralama yapılmaz. Tutarlılığı sağlamak için, sistemdeki her bloğa bir **geri işaretçi (back Pointer)** eklenir; örneğin, her veri bloğunun ait olduğu inode için bir referans vardır. Bir dosyaya erişirken, dosya sistemi, ileri işaretçinin (örneğin, inode veya doğrudan bloktaki adres) dosyaya geri başvuran bir bloğa işaret edip etmediğini kontrol ederek dosyanın tutarlı olup olmadığını belirleyebilir. Eğer öyleyse, her şey güvenli bir şekilde diske ulaşmış olmalı ve bu nedenle dosya tutarlıdır; değilse, dosya tutarsızdır ve bir hata döndürür. Dosya sistemine geri işaretçiler ekleyerek, yeni bir tembel kilitlenme tutarlılığı biçimi elde edilebilir [C+12].

Son olarak, bir günlük protokolünün disk yazmalarının tamamlanması için beklemesi gereken süreyi azaltan teknikleri de araştırdık. **İyimser kilitlenme tutarlılığı (optimistic crash consistency)** [C+13] olarak adlandırılan bu yeni yaklaşım, **işlem sağlama toplamının (transaction checksum)** genelleştirilmiş bir biçimini[P+05] kullanarak diske münkün olduğunca çok yazma işlemi gerçekleştirir ve ortaya çıkmaları durumunda tutarsızlıkları tespit etmek için birkaç başka teknik içerir. Bazı iş yükleri için bu iyimser teknikler performansı büyüklük sırasına göre artırabilir. Ancak, gerçekten iyi çalışması için, biraz farklı bir disk arayüzü gereklidir [C+13].

## Özet ( Summary )

Kilitlenme tutarlılığı sorununu ortaya koyduk ve bu soruna saldırmak için çeşitli yaklaşımları tartıştık. Bir dosya sistemi kontrolcüsü oluşturmanın eski yaklaşımı işe yarar, ancak modern sistemlerde kurtarılmayacak kadar yavaştır. Bu nedenle, birçok dosya sistemi artık günlük kaydı kullanmaktadır. Günlük kaydı, kurtarma süresini O’dan (disk hacminin boyutu) O’ya (günlük boyutu) düşürür, böylece bir kilitlenme ve yeniden başlatmadan sonra kurtarma işlemini önemli ölçüde hızlandırır. Bu nedenle, birçok modern dosya sistemi günlük kaydını kullanır. Günlük kaydının birçok farklı biçimde gelebileceğini de gördük; en sık kullanılan, hem dosya sistemi meta verileri hem de kullanıcı verileri için makul tutarlılık garantilerini korurken, günlüğe giden trafik miktarını azaltan sıralı meta veri günlük kaydıdır. Sonunda, kullanıcı verileri üzerinde güçlü garantiler muhtemelen sağlanması gereken en önemli şeylerden biridir; garip bir şekilde, son araştırmaların gösterdiği gibi, bu alan devam eden bir çalışma olmaya devam ediyor. [P+14].

# References

[B07] “ZFS: The Last Word in File Systems” by Jeff Bonwick and Bill Moore. Available online: [http://www.ostep.org/Citations/zfs\_last.pdf.](http://www.ostep.org/Citations/zfs_last.pdf) *ZFS uses copy-on-write and journal- ing, actually, as in some cases, logging writes to disk will perform better.*

[C+12] “Consistency Without Ordering” by Vijay Chidambaram, Tushar Sharma, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. FAST ’12, San Jose, California. *A recent paper of ours about a new form of crash consistency based on back pointers. Read it for the exciting details!*

[C+13] “Optimistic Crash Consistency” by Vijay Chidambaram, Thanu S. Pillai, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau . SOSP ’13, Nemacolin Woodlands Resort, PA, November 2013. *Our work on a more optimistic and higher performance journaling protocol. For workloads that call fsync() a lot, performance can be greatly improved.*

[GP94] “Metadata Update Performance in File Systems” by Gregory R. Ganger and Yale N. Patt. OSDI ’94. *A clever paper about using careful ordering of writes as the main way to achieve consistency. Implemented later in BSD-based systems.*

[G+08] “SQCK: A Declarative File System Checker” by Haryadi S. Gunawi, Abhishek Ra- jimwale, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. OSDI ’08, San Diego, Califor- nia. *Our own paper on a new and better way to build a file system checker using SQL queries. We also show some problems with the existing checker, finding numerous bugs and odd behaviors, a direct result of the complexity of fsck.*

[H87] “Reimplementing the Cedar File System Using Logging and Group Commit” by Robert Hagmann. SOSP ’87, Austin, Texas, November 1987. *The first work (that we know of) that applied write-ahead logging (a.k.a. journaling) to a file system.*

[M+13] “ffsck: The Fast File System Checker” by Ao Ma, Chris Dragga, Andrea C. Arpaci- Dusseau, Remzi H. Arpaci-Dusseau. FAST ’13, San Jose, California, February 2013. *A recent paper of ours detailing how to make fsck an order of magnitude faster. Some of the ideas have already been incorporated into the BSD file system checker [MK96] and are deployed today.*

[MK96] “Fsck – The UNIX File System Check Program” by Marshall Kirk McKusick and T. J. Kowalski. Revised in 1996. *Describes the first comprehensive file-system checking tool, the epony- mous fsck. Written by some of the same people who brought you FFS.*

[MJLF84] “A Fast File System for UNIX” by Marshall K. McKusick, William N. Joy, Sam J. Leffler, Robert S. Fabry. ACM Transactions on Computing Systems, Volume 2:3, August 1984. *You already know enough about FFS, right? But come on, it is OK to re-reference important papers.*

[P+14] “All File Systems Are Not Created Equal: On the Complexity of Crafting Crash-Consistent Applications” by Thanumalayan Sankaranarayana Pillai, Vijay Chidambaram, Ramnatthan Alagappan, Samer Al-Kiswany, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. OSDI ’14, Broomfield, Colorado, October 2014. *A paper in which we study what file systems guarantee after crashes, and show that applications expect something different, leading to all sorts of interesting problems.*

[P+05] “IRON File Systems” by Vijayan Prabhakaran, Lakshmi N. Bairavasundaram, Nitin Agrawal, Haryadi S. Gunawi, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. SOSP ’05, Brighton, England, October 2005. *A paper mostly focused on studying how file systems react to disk failures. Towards the end, we introduce a transaction checksum to speed up logging, which was eventually adopted into Linux ext4.*

[PAA05] “Analysis and Evolution of Journaling File Systems” by Vijayan Prabhakaran, Andrea

C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. USENIX ’05, Anaheim, California, April 2005.

*An early paper we wrote analyzing how journaling file systems work.*

[R+11] “Coerced Cache Eviction and Discreet-Mode Journaling” by Abhishek Rajimwale, Vijay Chidambaram, Deepak Ramamurthi, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. DSN ’11, Hong Kong, China, June 2011. *Our own paper on the problem of disks that buffer writes in a memory cache instead of forcing them to disk, even when explicitly told not to do that! Our solution to overcome this problem: if you want A to be written to disk before B, first write A, then send a lot of “dummy” writes to disk, hopefully causing A to be forced to disk to make room for them in the cache. A neat if impractical solution.*

[T98] “Journaling the Linux ext2fs File System” by Stephen C. Tweedie. The Fourth Annual Linux Expo, May 1998. *Tweedie did much of the heavy lifting in adding journaling to the Linux ext2 file system; the result, not surprisingly, is called ext3. Some nice design decisions include the strong focus on backwards compatibility, e.g., you can just add a journaling file to an existing ext2 file system and then mount it as an ext3 file system.*

[T00] “EXT3, Journaling Filesystem” by Stephen Tweedie. Talk at the Ottawa Linux Sympo- sium, July 2000. olstrans.sourceforge.net/release/OLS2000-ext3/OLS2000-ext3.html *A tran- script of a talk given by Tweedie on ext3.*

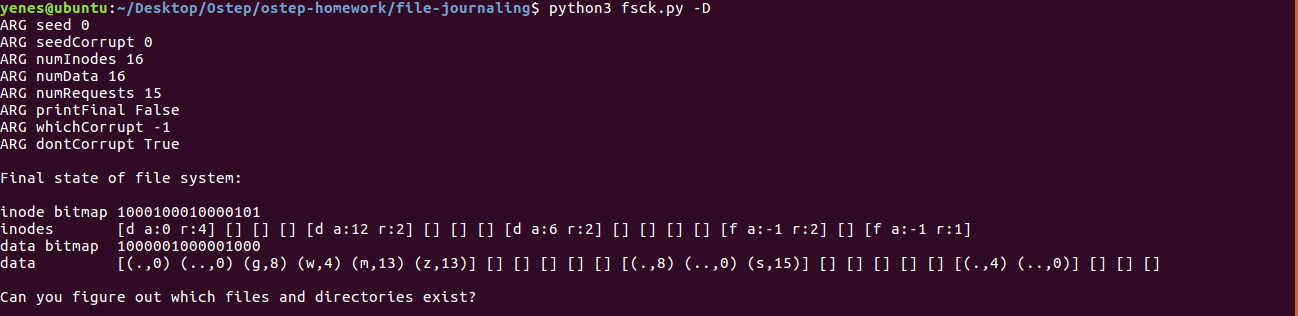
[T01] “The Linux ext2 File System” by Theodore Ts’o, June, 2001.. Available online here: [http://e2fsprogs.sourceforge.net/ext2.html.](http://e2fsprogs.sourceforge.net/ext2.html) *A simple Linux file system based on the ideas found in FFS. For a while it was quite heavily used; now it is really just in the kernel as an example of a simple file system.*

# Ödev (Similasyon)

Bu bölümde, dosya sistemi bozulmalarının nasıl tespit edilebileciğini (ve potansiyel olarak onarılmış)daha iyi anlamak için kullanabileceğiniz basit bir simülatör olan fsck.py tanıtılmaktadır. Simülatörün nasıl çalıştırılacağı hakkında ayrıntılar için lütfen ilgili README’ye bakın.

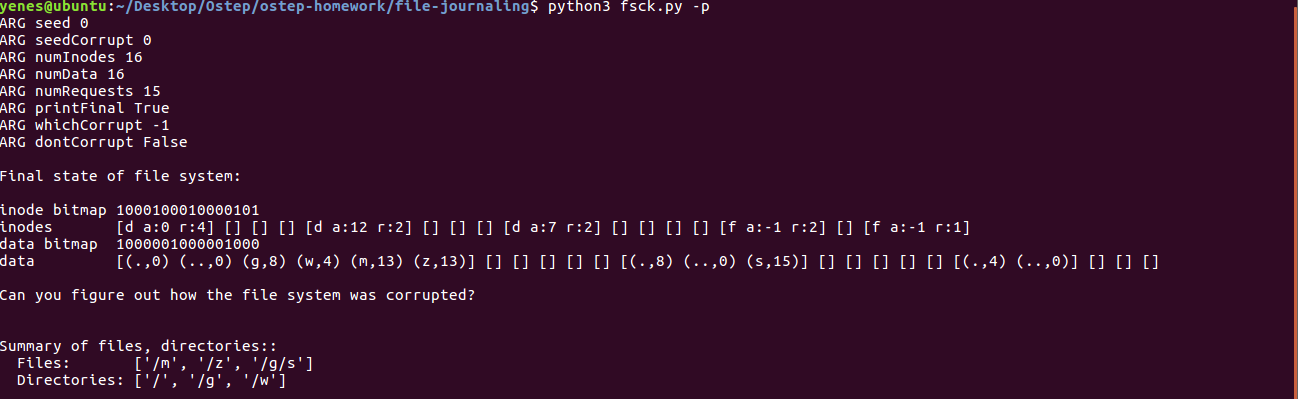
### Sorular

1. İlk olarak, fsck.py –D dosyasını çalıştırın; bu bayrak herhangi bir tutarsızlığı kapatır ve böylece rastgele bir dosya sistemi oluşturmak için kullanabilir ve orada hangi dosya ve dizinlerin olduğunu belirleyip bakabilirsiniz. Öyleyse, devam edin ve bunu yapın! Haklı olup olmadığınızı görmek için -p bayrağını kullanın. Tohumu (-s) 1,2 ve 3 gibi farklı değerlere ayarlayarak bunu rastgele oluşturulmuş birkaç farklı dosya sistemi için deneyin.

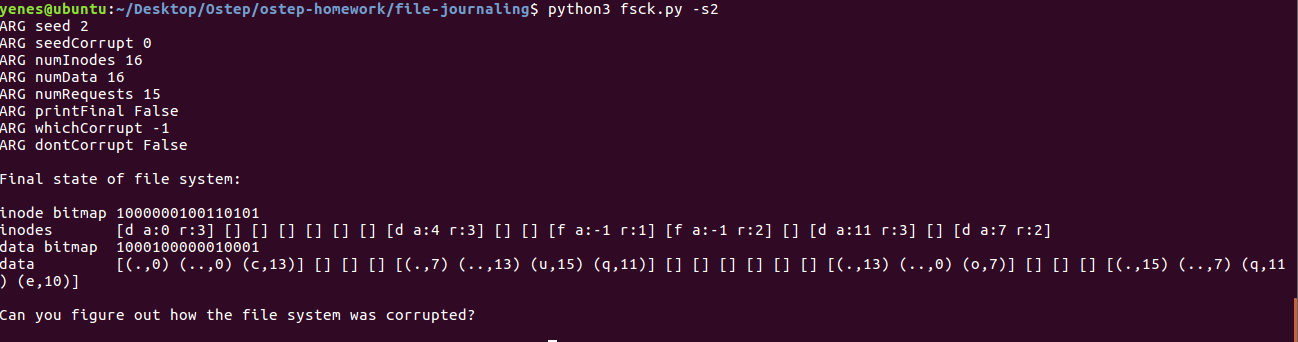
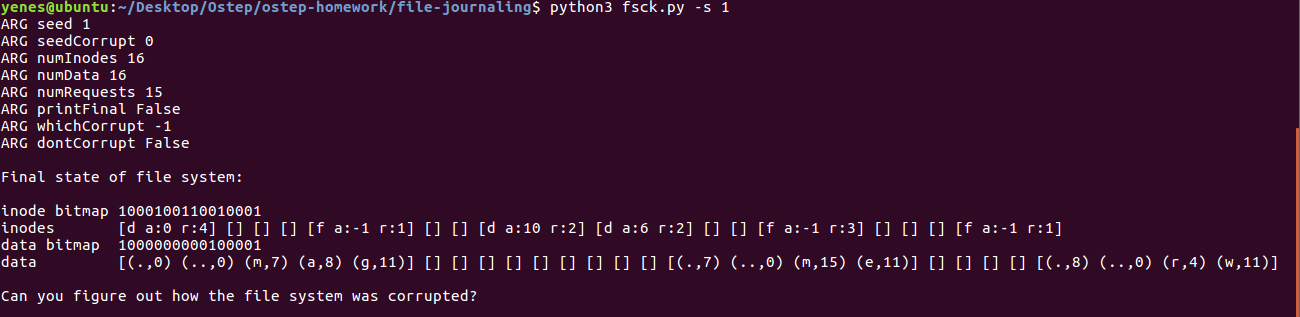


-D bayrağı dosya sisteminin bozulmalarını kapatır ve böylelikle bozuk olan dosya sistemiyle karşılaştırdığımızda bozulmayı tanımlayabiliriz. Bozulmanın olmadığını dontCorrupt True olduğunda anlıyoruz eğer False ise dosya sistemi bozulmaya uğramıştır.

-p bayrağı son dosya veya dizini yazdırır. Burada tohum 0 da bir bozulma vardır ve -D bayrağı ile oluşturduğumuz dosya sistemiyle karşılaştırdığımızda bozulma inode 8’in ölü blok 7’yi işaret ettiğini görüyoruz.

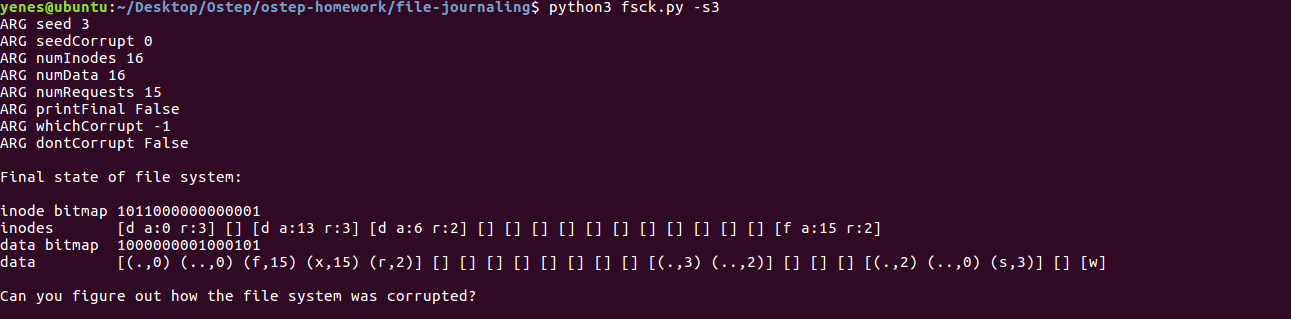


-s bayrağı ile rastgele dosya sistemlerinin tohumlarını kullanıyoruz ve böylelikle aynı sorunu farklı dosya sistemlerinde görebiliyoruz. Örneğin, yukarıda seed 0 idi ve inode 8 ölü blok 7’yi işaret ediyordu fakat –s 1 yaptığımızda yani tohumu 1 olarak ayarladığımızda inode 8 ölü blok 6’yı işaret ediyor.

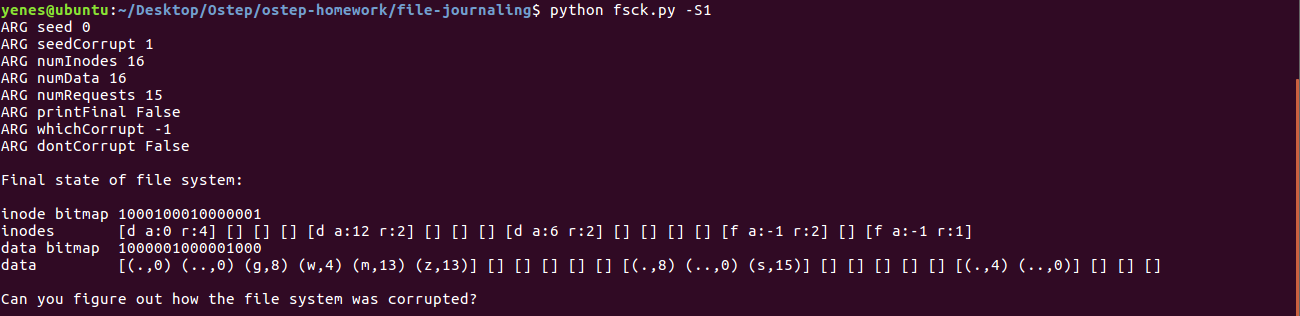


-s 2 ile tohumu 2 olarak ayarladığımızda inode 15 ölü blok 7’yi işaret ediyor.

-s 3 ile tohumu 3 olarak ayarladığımızda ise inode 4 ölü blok 6’yı işaret ediyor. Gördüğünüz gibi farklı dosya sistemlerindeki ortak sorun inode’larımızın ölü bloğu işaret etmesiydi bunun sebebi ise seed’in altında yer alan seedCorrupttur. SeedCorrupt Bozulmanın tohumudur sonraki sorularımızda bunu kullanacağız ve bozulmaların nasıl değiştiğini göreceksiniz.



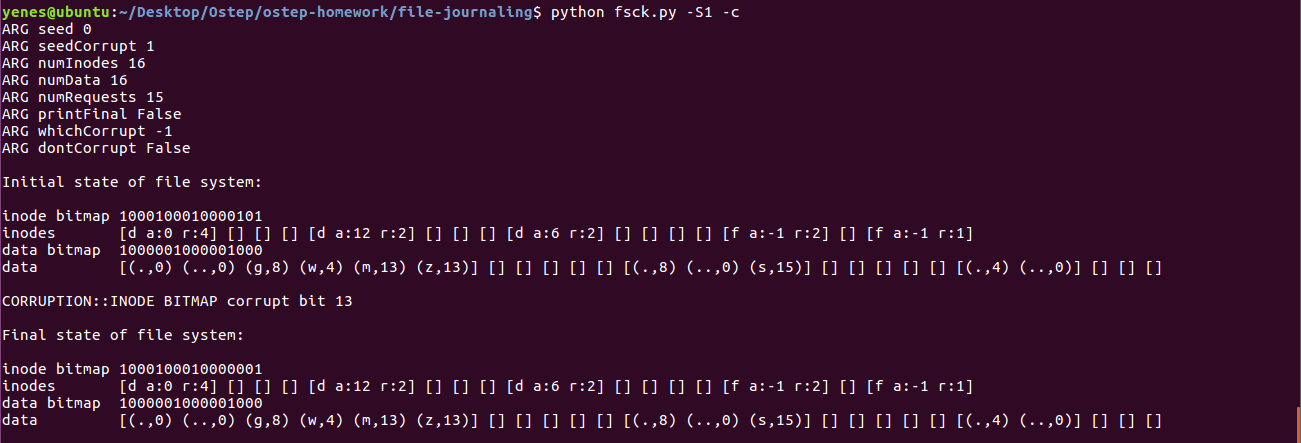
1. Şimdi, bir tutarsızlığı tanıtalım. Başlamak için fsck.py -S 1 dosyasını çalıştırın. Hangi tutarsızlığı ortaya çıktığını görebiliyor musunuz? Gerçek bir dosya sistemi onarım aracında bunu nasıl düzeltirdiniz? Haklı olup olmadığınızı kontrol etmek için -c kullanın.



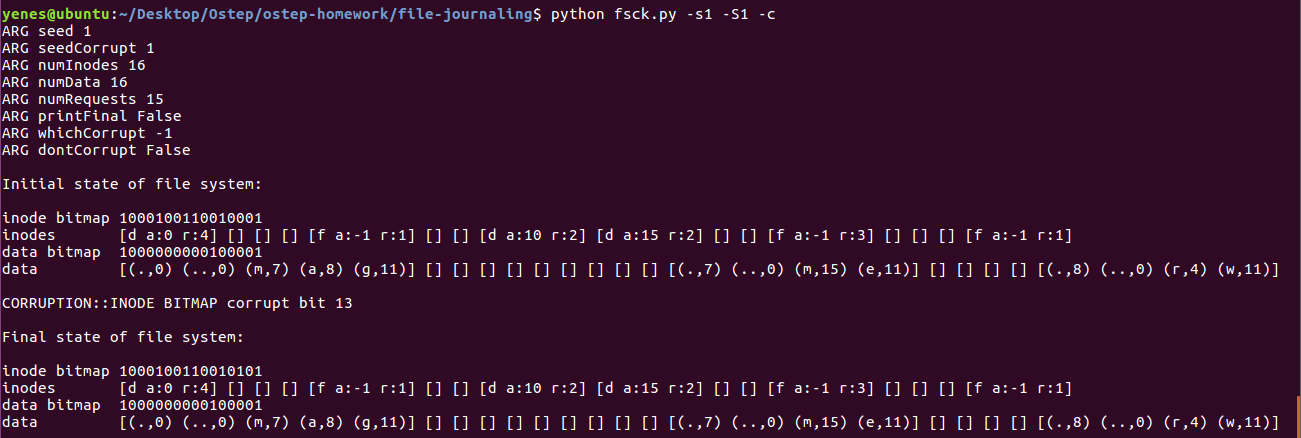
-S1 bayrağını çalıştırdığımızda tohumu 0 olan dosya sisteminin bozulma tohumu 1’ini görüyoruz. Öncelikle –D bayrağı ile Bozulmamış haline bakıp karşılaştıralım.



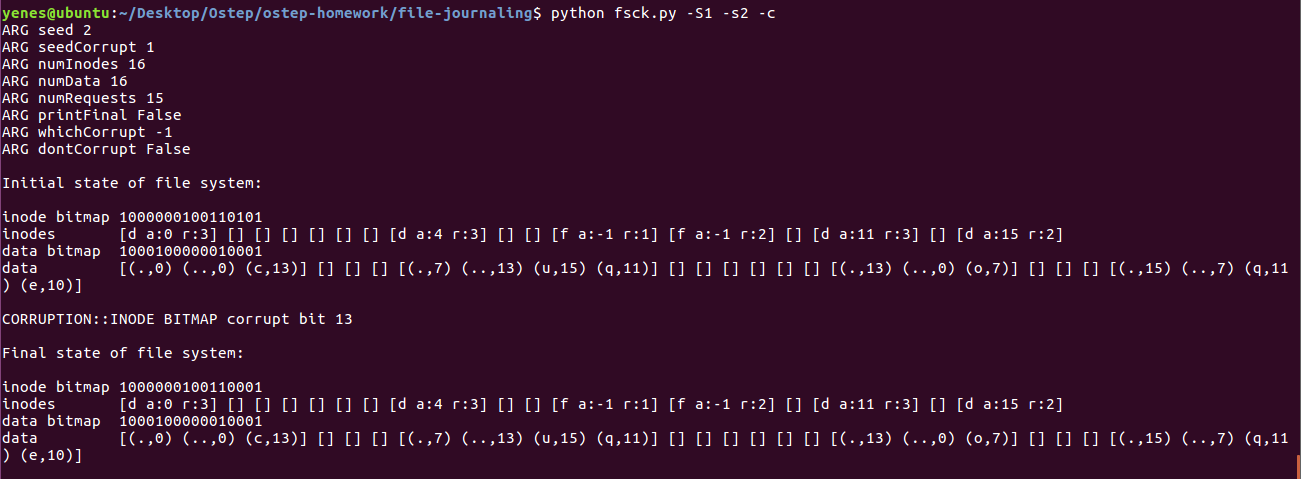
Karşılaştırdığımızda bozulmamızın inode kısmında tahsis edilmiş 13. İnode’un, inode bitmapinde işaretlenmemiş olarak görüyoruz. İnode’u Bitmap üzerinde işaretlersek bozulmamız çözülür. –c bayrağı ile bulduğu-muz bozulmanın doğru olup olmadığını kontrol edelim.



-c bayrağı ile çalıştırdığımızda bozulmanın inode bitmap’inde 13. bitte olduğunu söylüyor yani cevabımız doğru. Seed değiştirerek farklı dosya sistemlerinde aynı hatanın nasıl olduğuna bakalım.



Tohumu –s1 bayrağı ile değiştirdiğimizde tohumu 1 olan dosya sistemimiz.

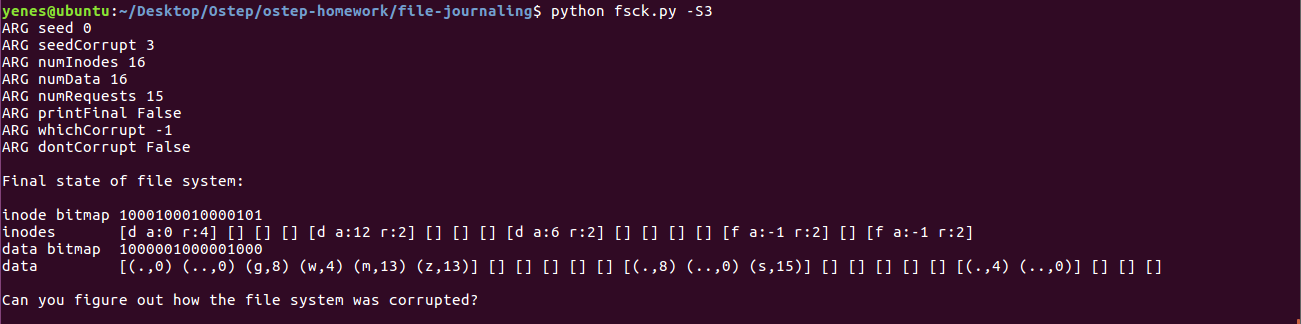


Tohumu –s2 bayrağı ile değiştirdiğimizde tohumu 2 olan dosya sistemimiz.

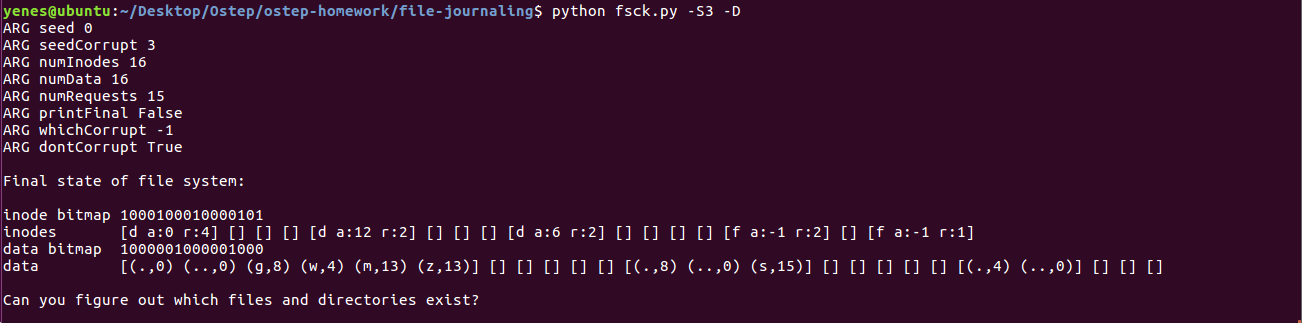


Tohumu –s3 bayrağı ile değiştirdiğimizde tohumu 3 olan dosya sistemimiz. Gördüğümüz üzere farklı dosya sistemlerinde aynı bozulmayı yani inode bitmapinde 13. Bitin bozulduğunu görüyoruz.

1. Tohumu -S 3 veya -S 19 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Cevabınızı kontrol etmek için -c kullanın. Bu iki durumda farklı olan nedir?



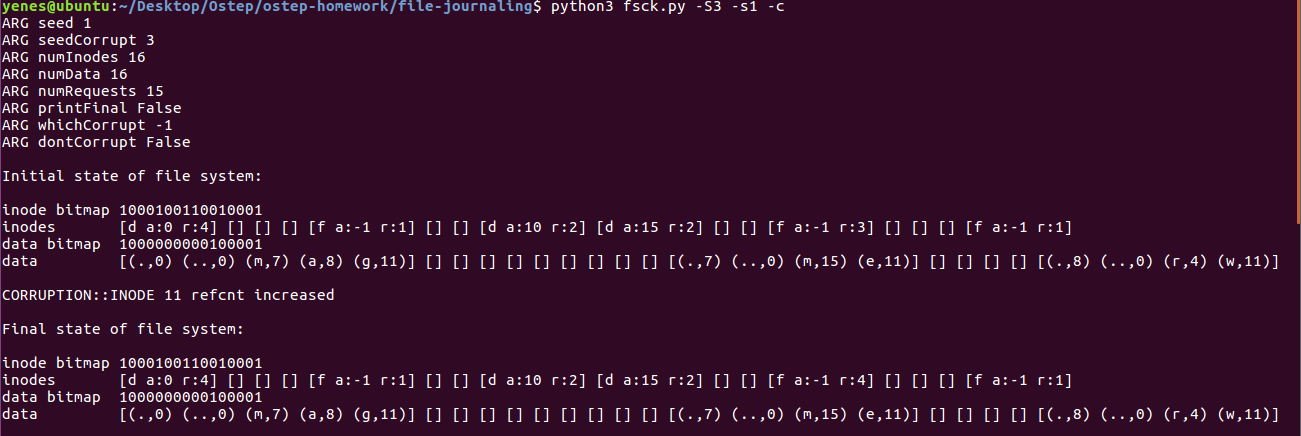
-S3 bayrağını çalıştırdığımızda tohumu 0 olan dosya sisteminin bozulma tohumu 3’ini görüyoruz. –D bayrağı kullanarak bozulmaya uğramamış haliyle karşılaştırma yapıp bozulmayı bulalım.



Gördüğümüz üzere 15. İnode’umuzun dosya referans sayısı bir artmış. Bu yüzden bir bozulma meydana gelmiş. Cevabımızı –c bayrağı ile doğru olup olmadığını kontrol edelim.



-c bayrağı ile çalıştırdığımızda inode 15’in referans sayısının arttığını söylüyor yani cevabımız doğru. Tohumları değiştirerek aynı bozulmayı farklı dosya sistemlerinde görelim.



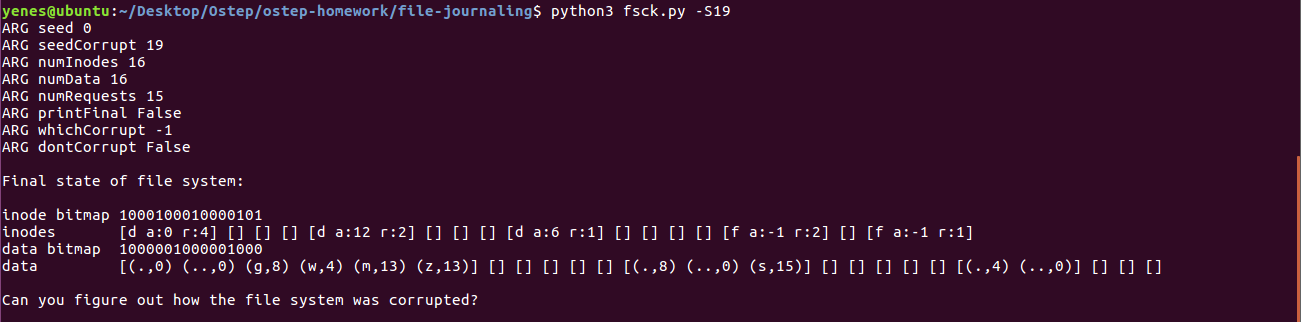
Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde inode 11’in referansının bir arttığını görüyoruz.



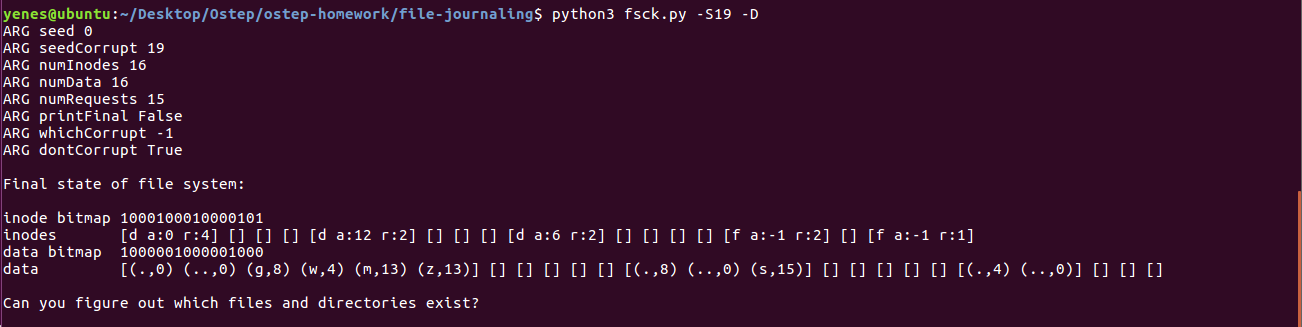
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde de inode 11’in referansının bir arttığını fakat dosya sistemlerinin farklı olduğunu açıkça görüyoruz.



Son olarak dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde inode 15’in referansının bir arttığını görüyoruz. –S3 bayrağı ile oluşturulan bozulmalarının ortak özelliği inode’lardan herhangi birinin referans numarasının bir artmasıdır. –s ile –S bayrağının farklı olduğunu unutmamak gerek.



-S19 bayrağını çalıştırdığımızda tohumu 0 olan dosya sisteminin bozulma tohumu 19’u görüyoruz. –D bayrağı kullanarak bozulmaya uğramamış haliyle karşılaştırma yapıp bozulmayı bulalım.



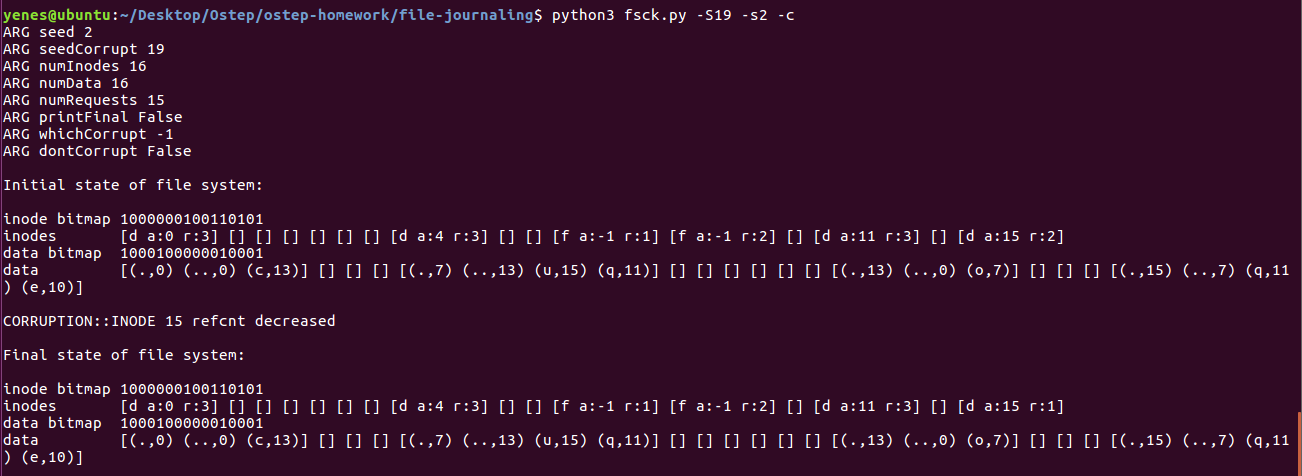
Burada da önceki bozulmamızın tam tersini görüyoruz. Bozulmamızın sebebi inode 8’in referans sayısının bir azalmasıdır. –c bayrağı ile cevabımızın doğruluğunu kontrol edelim.



-c bayrağı ile çalıştırdığımızda inode 8’in referans sayısının bir azaldığını söylüyor ve böylelikle cevabımızın doğru olduğunu görüyoruz. Dosya sistemi tohumunu değiştirerek diğer dosya sistemlerinde bozulmalarına bakalım.



Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde inode 8’in referansının bir azaldığını görüyoruz.



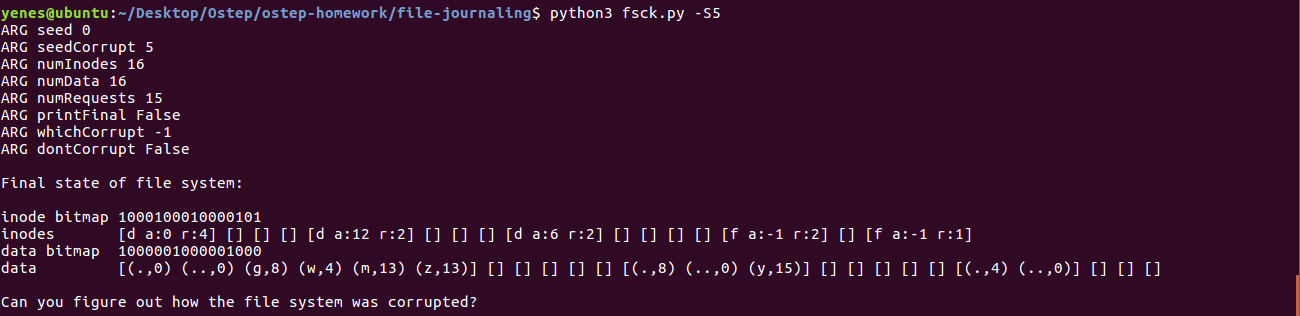
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde inode 15’in referansının bir azaldığını görüyoruz.



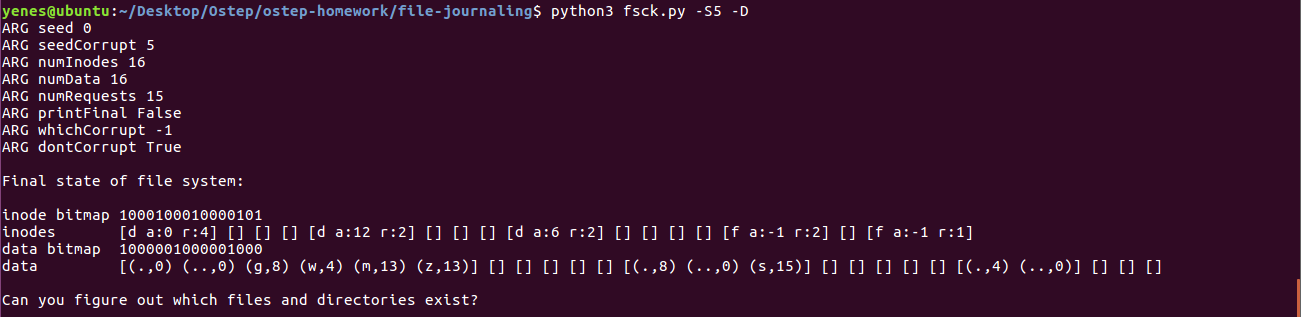
Son olarak dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde inode 3’ün referans sayısının bir azaldığını görüyoruz. –S19 bayrağı ile oluşturulan bozulmalarının ortak özelliği inode’lardan herhangi birinin referans numarasının bir azalmasıdır.

Sonuç olarak –S3 bayrağı referans numarasının bir artmasını –S19 bayrağı ise referans numarasının bir azaldığı bozulmaları gösterir.

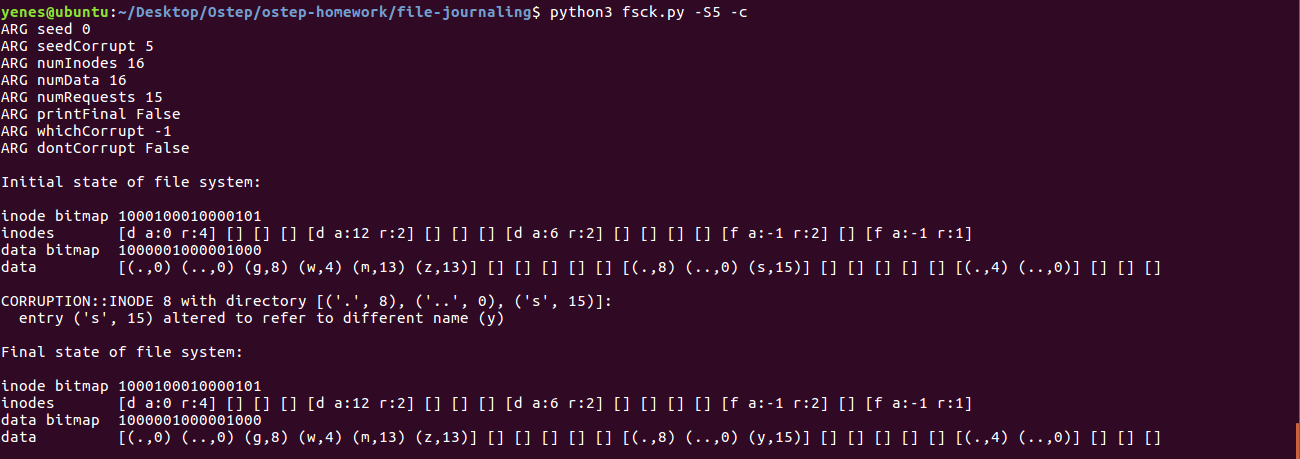
1. Tohumu -S 5 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Bu sorunu otomatik bir şekilde düzeltmek ne kadar zor olurdu? Cevabınızı kontrol etmek için -c kullanın. Ardından, -S 38 ile benzer bir bozulma ekleyin; bunu tespit etmek daha mı zor/mümkün? Son olarak, -S 642 kullanın; bu bozulma tespit edilebilir mi? Eğer öyleyse, dosya sistemini nasıl düzeltirsiniz?



Tohumu -S5 bayrağı olarak değiştirdiğimizde tohumu 0 olan dosya sisteminin bozulma tohumu 5’ini görmekteyiz. –D bayrağını kullanarak bozulmaya uğramayan haliyle karşılaştıralım.

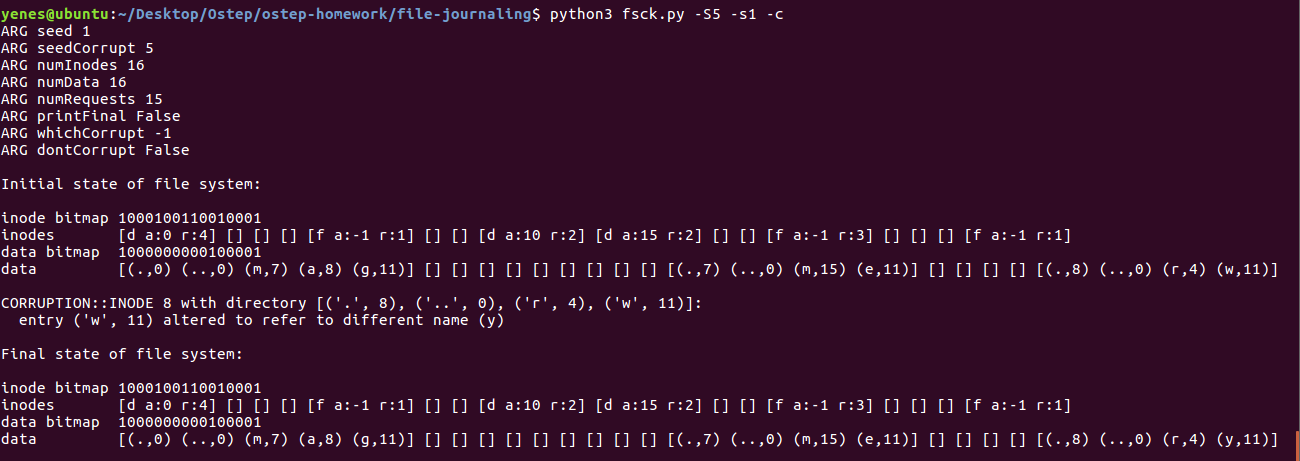


Karşılaştırdığımızda 8. İnode’un işaret ettiği data bölgesindeki 6. Adres bloğundaki (s,15) verisi (y,15) olarak değiştiği için bozulma meydana gelmektedir. Bozulmalar data bölgesinde olduğu için düzeltmek zor olmaz. Cevabımızı –c bayrağını çalıştırarak doğru olup olmadığını kontrol edelim.

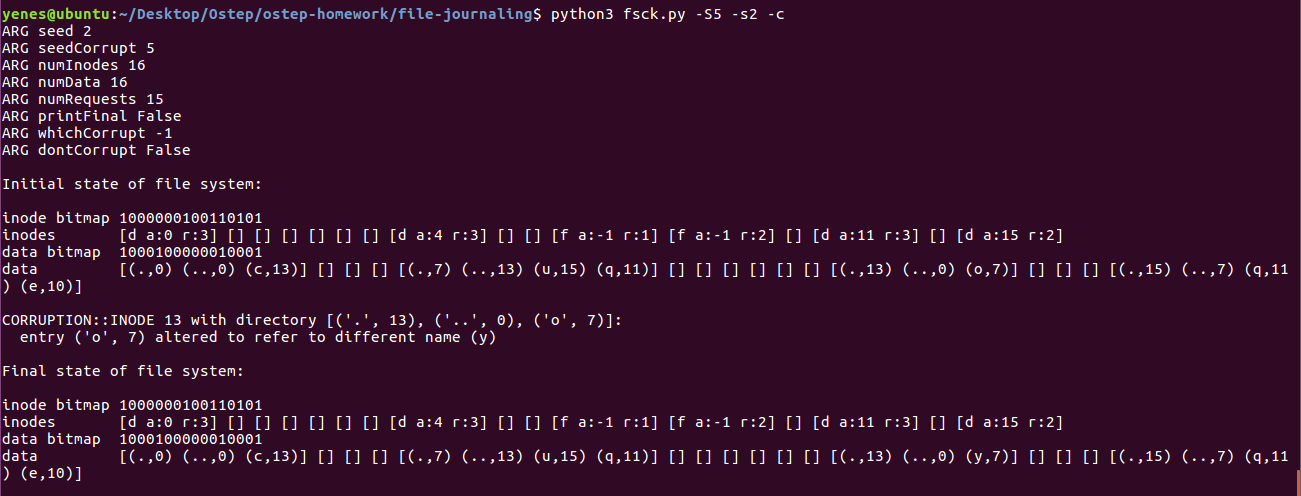


-c bayrağı ile çalıştırdığımızda bize bozulmanın sebebinin inode 8’in dizininde giriş ('s', 15) farklı bir isme (y) referans için değiştirildiğini söylüyor yani cevabımız doğrudur.

–s ile dosya sistemi tohumlarımızı değiştirip farklı dosya sistemlerinde bozulmamıza tekrar bakalım.



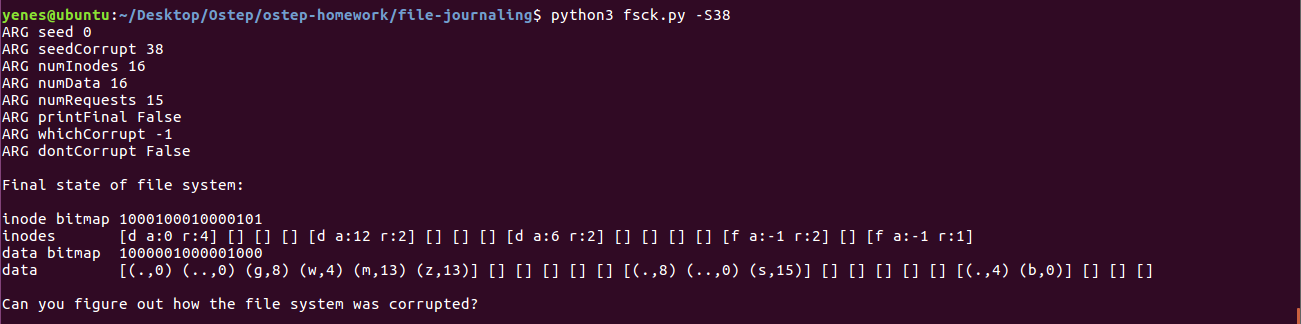
Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde 8. İnode’un işaret ettiği data bölgesindeki 15. adres bloğundaki (w,11) verisi (y,11) olarak değiştiğini görüyoruz.



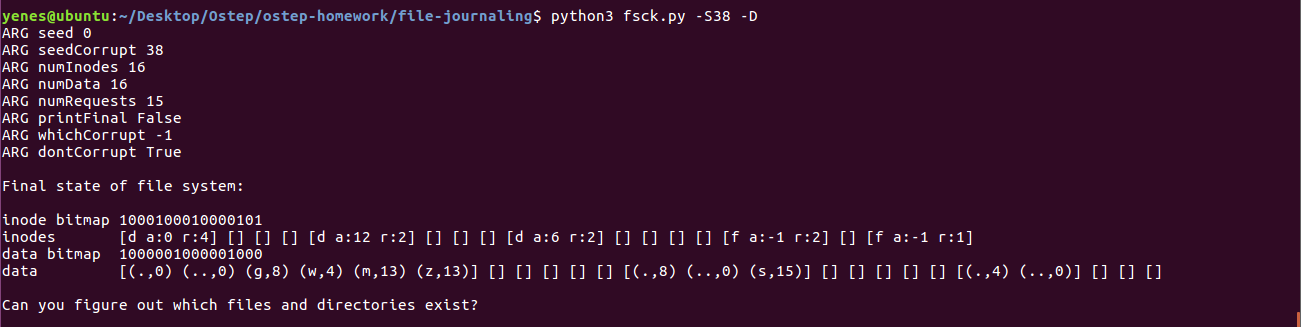
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde 13. İnode’un işaret ettiği data bölgesindeki 11. adres bloğundaki (o,7) verisi (y,7) olarak değiştiğini görüyoruz.



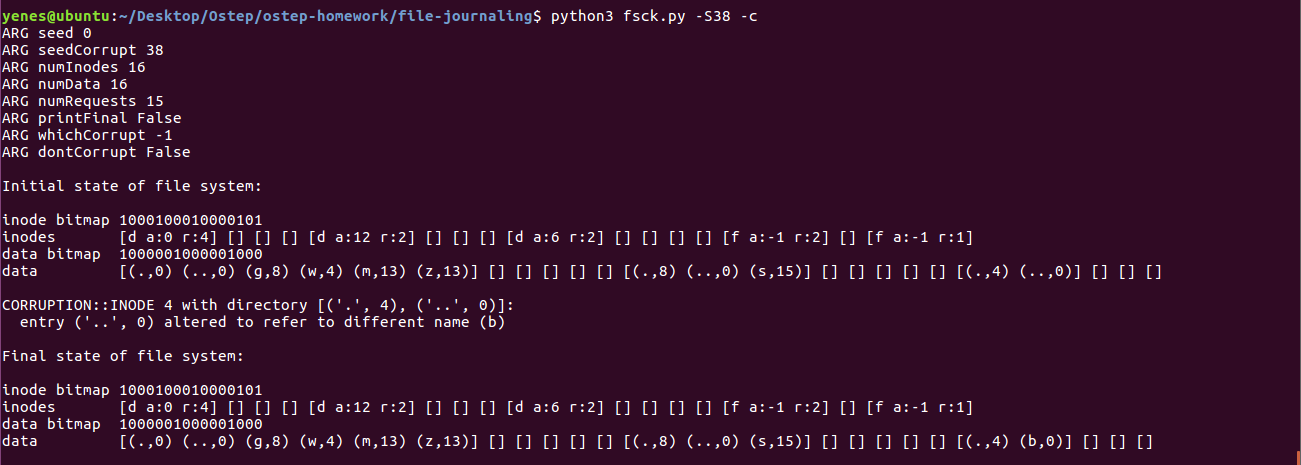
Son olarak Dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde 3. İnode’un işaret ettiği data bölgesindeki 13. adres bloğundaki (..,2) verisi (y,2) olarak değiştiğini görüyoruz.



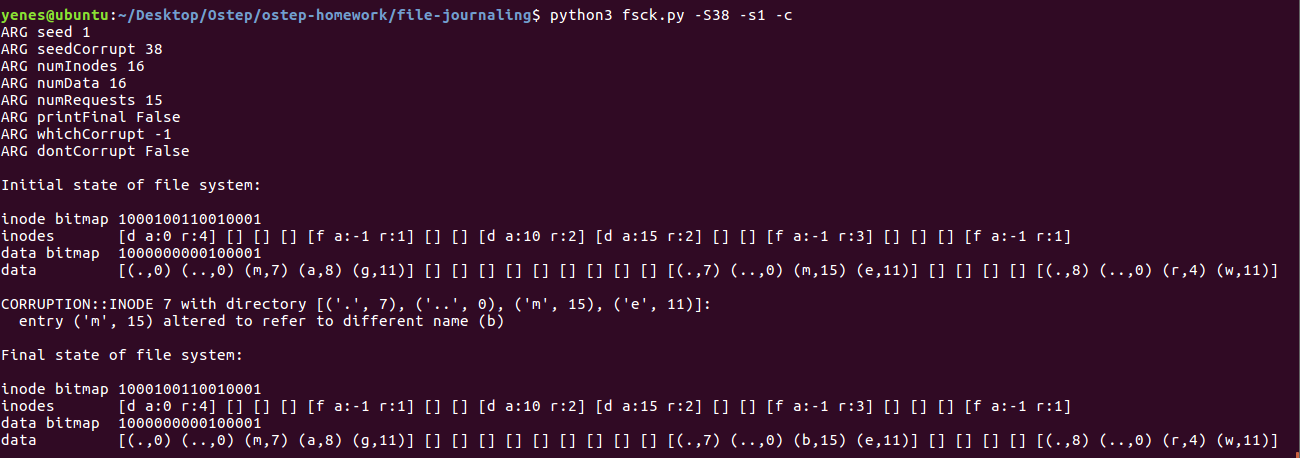
-S38 bayrağı ile farklı bir bozulmayı görüyoruz. Bozulmayı tespit etmek için –D bayrağı ile bozulmamış halini çalıştıralım ve karşılaştıralım.



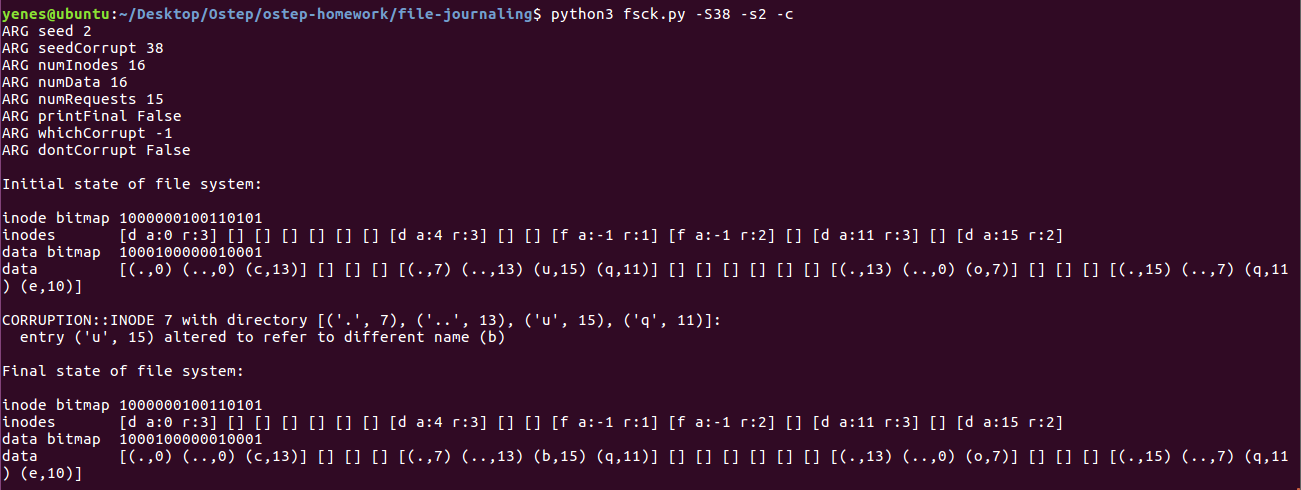
Karşılaştırdığımızda 4. İnode’un işaret ettiği data bölgesindeki 12. Adres bloğundaki (..,0) verisi (b,0) olarak değiştiği için bozulma meydana gelmektedir. Bu bozulmamız da data bölgesinde gerçekleşmiştir ve bozulmamızı tespit etmek kolaydır. Cevabımızı –c bayrağını çalıştırarak doğru olup olmadığını kontrol edelim.



-c bayrağı ile çalıştırdığımızda bize bozulmanın sebebinin inode 4’ün dizininde giriş ('..', 0) farklı bir isme (b) referans için değiştirildiğini söylüyor yani cevabımız doğrudur. –s ile dosya sistemi tohumlarımızı değiştirip farklı dosya sistemlerinde bozulmamıza tekrar bakalım.



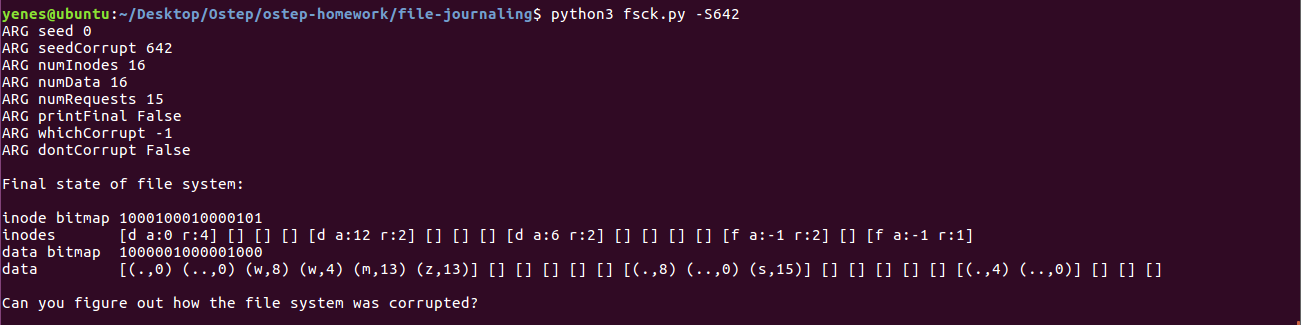
Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde 7. İnode’un işaret ettiği data bölgesindeki 10. adres bloğundaki (m,15) verisi (b,15) olarak değiştiğini görüyoruz.



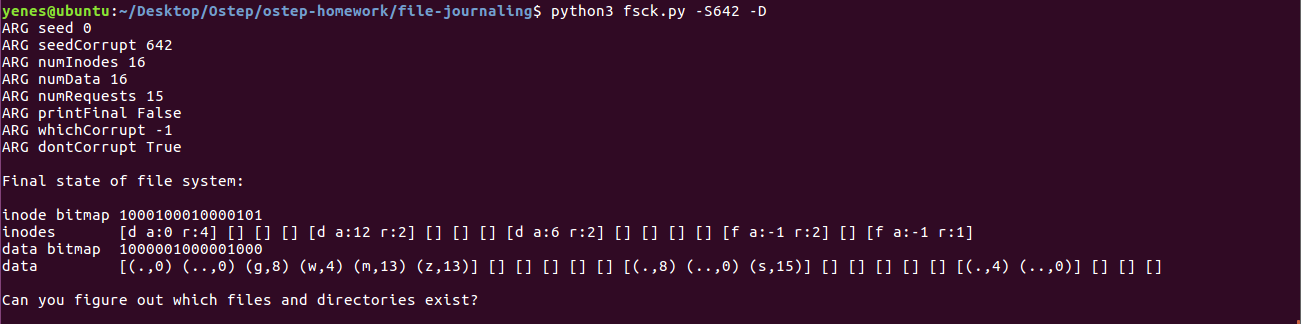
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde 7. İnode’un işaret ettiği data bölgesindeki 4. adres bloğundaki (u,15) verisi (b,15) olarak değiştiğini görüyoruz.



Son olarak Dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde 2. İnode’un işaret ettiği data bölgesindeki 13. adres bloğundaki (s,3) verisi (b,3) olarak değiştiğini görüyoruz.



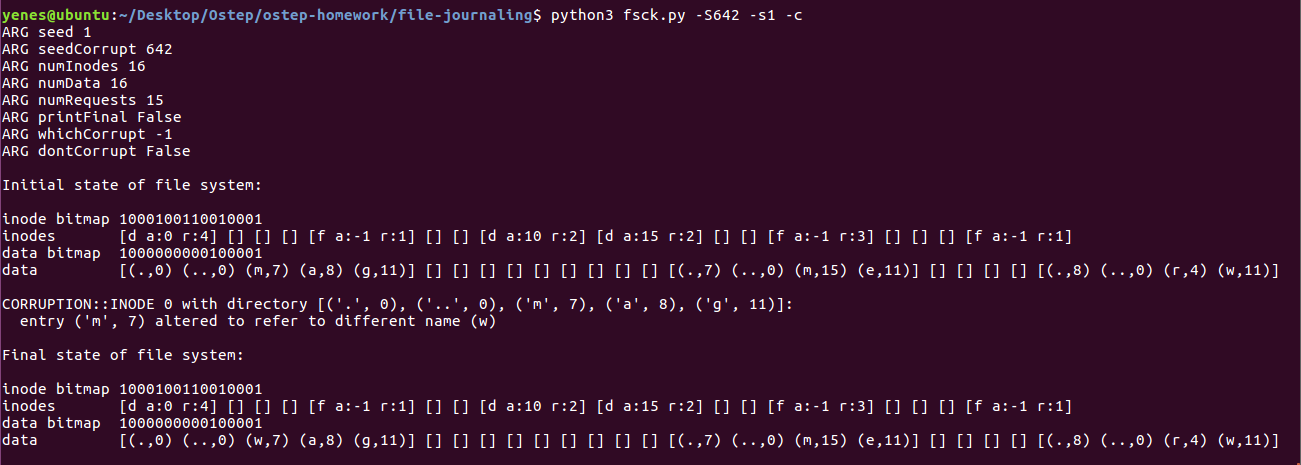
-S642 bayrağını çalıştırdığımızda yeni bir bozulma görüyoruz. Bozulmayı tespit etmek için yeniden –D bayrağı ile dosya sisteminin bozulmaya uğramamış halini çalıştırmamız gerekiyor.



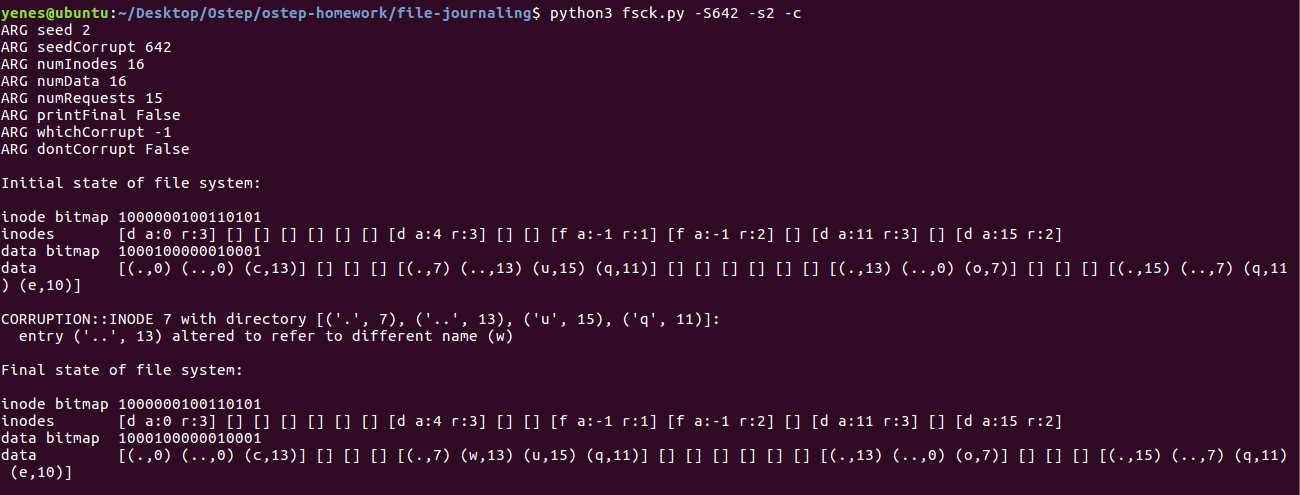
Çalıştırıp karşılaştırdığımızda 0. İnode’un işaret ettiği data bölgesindeki 0. Adres bloğundaki (g,8) verisi (w,8) olarak değiştiği için bozulma meydana gelmektedir. Bozulmamız data bölgesinde ve böylelikle kullanıcı tarafından değiştirilebilir, düzeltilebilir. Cevabımızı –c bayrağını çalıştırıp kontrol edebiliriz.



-c bayrağı ile çalıştırdığımızda bize bozulmanın sebebinin inode 0’ın dizininde giriş ('g', 8) farklı bir isme (w) referans için değiştirildiğini söylüyor yani cevabımız doğrudur. –s ile dosya sistemi tohumlarımızı değiştirip farklı dosya sistemlerinde bozulmamıza tekrar bakalım.



Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde 0. İnode’un işaret ettiği data bölgesindeki 0. adres bloğundaki (m,7) verisi (w,7) olarak değiştiğini görüyoruz.



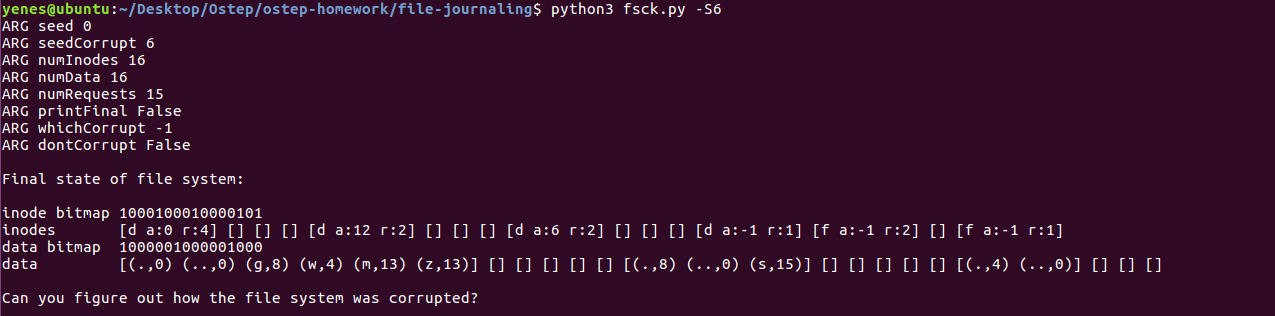
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde 7. İnode’un işaret ettiği data bölgesindeki 4. adres bloğundaki (..,13) verisi (w,13) olarak değiştiğini görüyoruz.



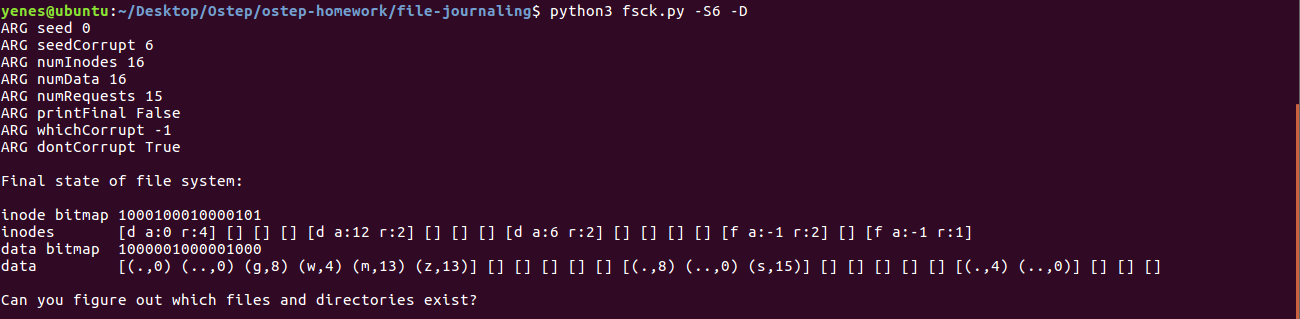
Son olarak Dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde 0. İnode’un işaret ettiği data bölgesindeki 0. adres bloğundaki (f,15) verisi (w,15) olarak değiştiğini görüyoruz.

4. Sorudaki sorulan tüm bozulmalar adres bloğundaki verinin değişmesiyle oluşmuştur.

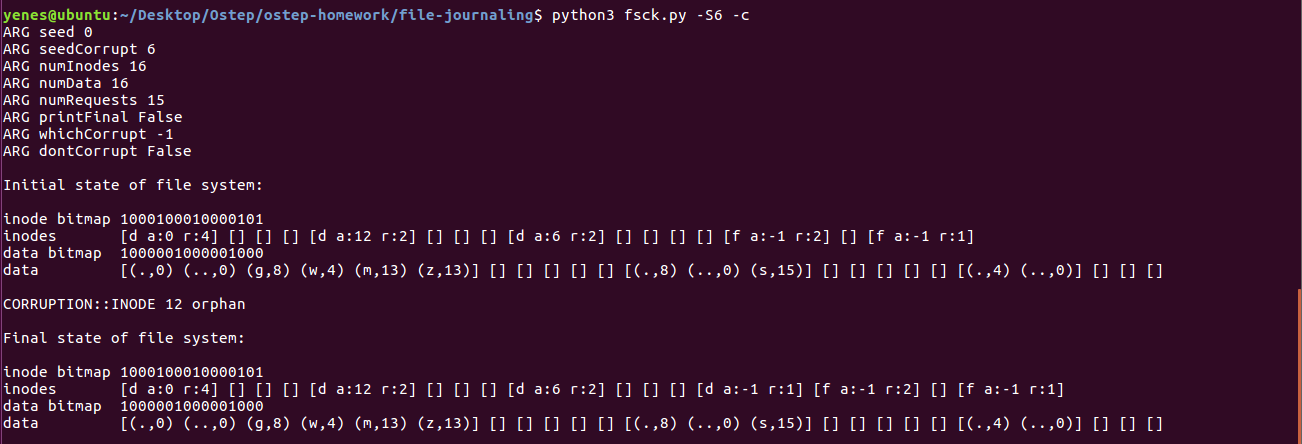
1. Tohumu -S 6 veya -S 13 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Cevabınızı kontrol etmek için -c kullanın. Bu iki durum arasındaki fark nedir? Onarım aracı böyle bir durumla karşılaştığında ne yapmalıdır?



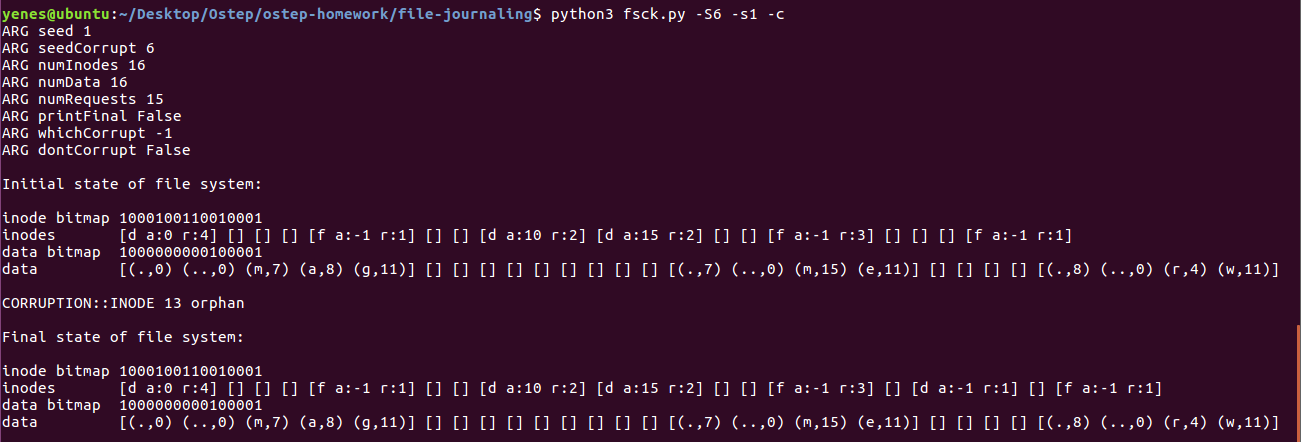
Tohumu –S6 bayrağı ile değiştirdiğimizde tohumu 0 olan dosya sisteminde bir bozulma meydana gelmektedir. Bozulmayı daha iyi görmek için –D bayrağını çalıştırıp bozulmaya uğramamış haliyle karşılaştıralım.



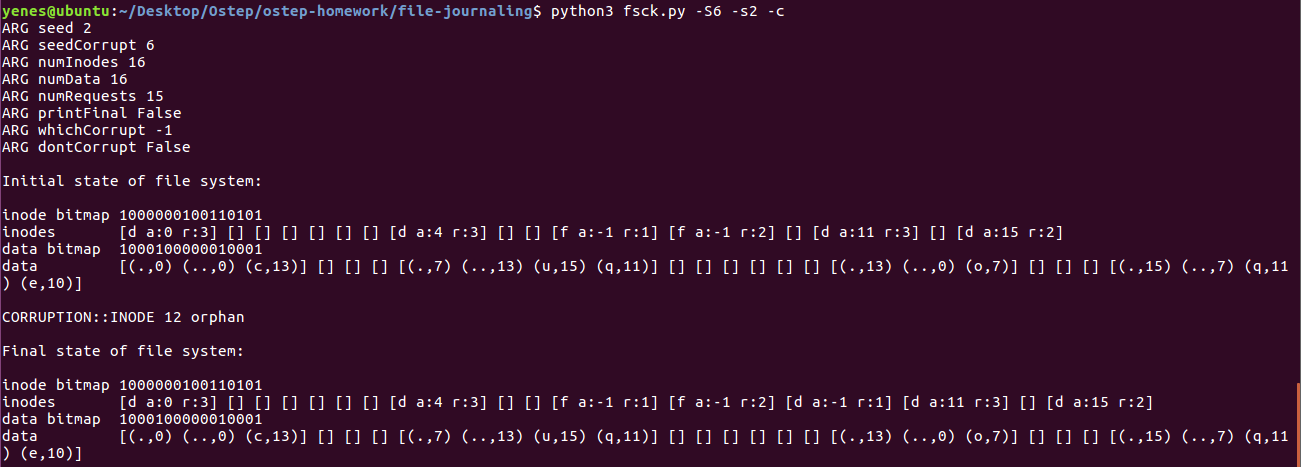
Karşılaştırdığımızda inode bitmap’inde tahsis edilmemesine rağmen 12. İnode doludur ve bozulma bu yüzdendir. Cevabımızı –c bayrağı kullanarak doğrulunu kontrol edelim.



-c bayrağını kullandığımızda inode 12’nin yetim olduğunu yani tahsis edilmediğini söylüyor. Burdan yola çıktığımızda cevabımız doğrudur. –s bayrakları ile farklı dosya sistemlerinde aynı bozulmayı görelim.



Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde inode bitmap’inde tahsis edilmemesine rağmen 13. İnode doludur.

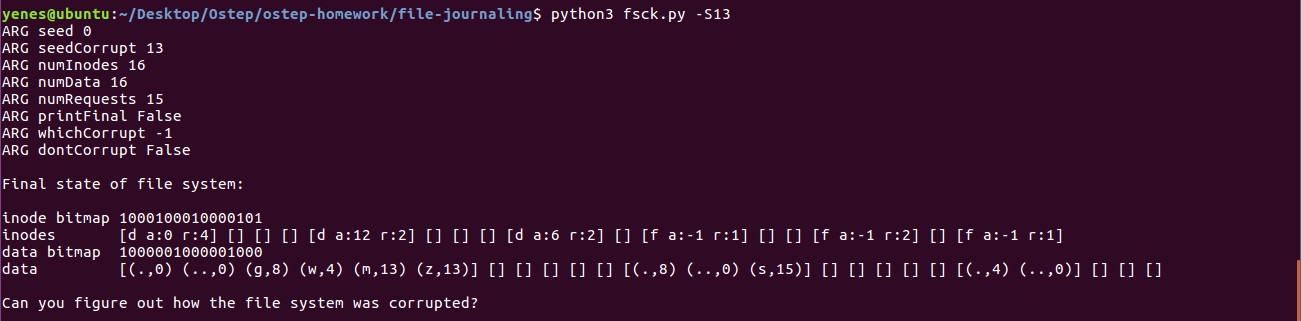


Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde inode bitmap’inde tahsis edilmemesine rağmen 12. İnode doludur.

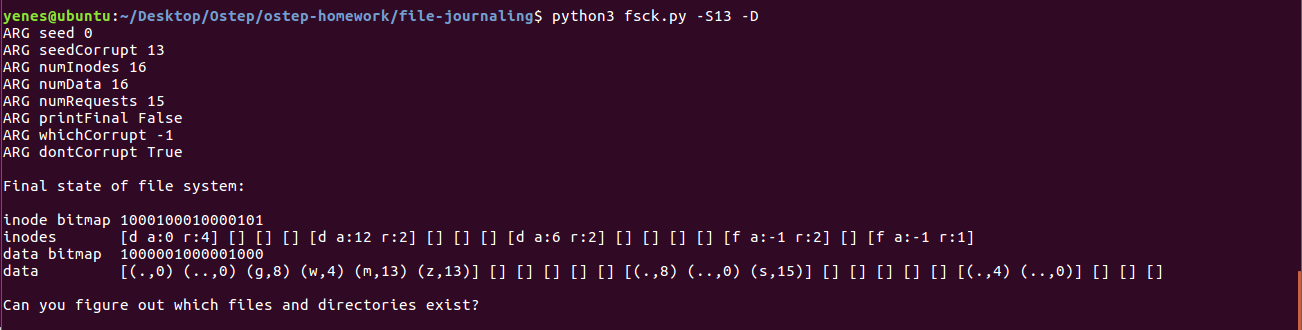


Son olarak Dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde de inode bitmap’inde tahsis edilmemesine rağmen 12. İnode doludur.

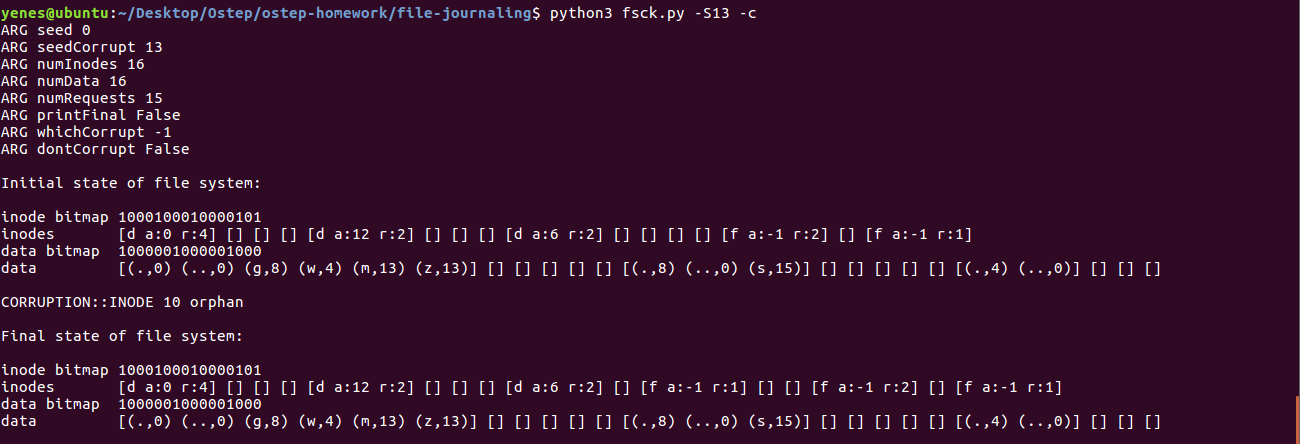
Bir de tohumu –S13 olarak değiştirelim ve bozulmaya bakalım.



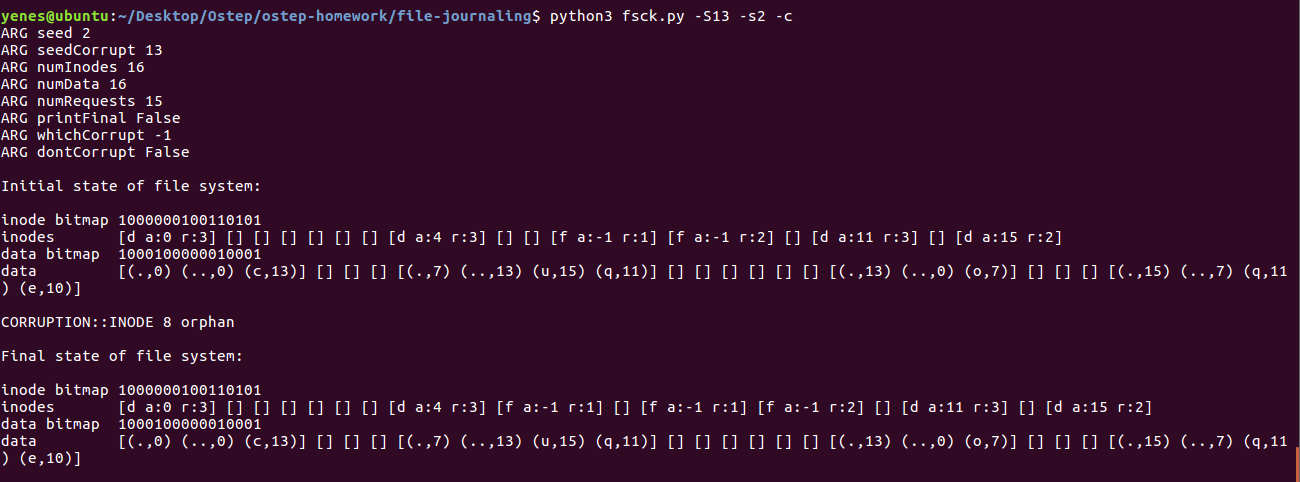
Tohumu 0 olan dosya sistemindeki bozulmayı saptamak için –D bloğu ile dosya sisteminin bozulmamış halini çalıştırıp karşılaştıralım.



Karşılaştırdığımızda inode bitmap’inde tahsis edilmemesine rağmen 10. İnode doludur ve bozulma bu yüzdendir. Önceki bozulmamızın aynısı diyebilirsiniz fakat aralarında bir fark var. Dikkatli bakarsak önceki yani –S6 bayrağı ile çalıştırdığımızda inode bitmap’de gözükmeyen ama için dolu olan inode [d a:-1 r:1] şeklindeydi. Burada baştaki d dosyanın türünün dizin olduğunu belirtir. Şuanki yani –S13 bayrağı ile çalıştırdığımızda inode bitmap’de gözükmeyen ama içi dolu olan inode [f a:-1 r:1] şeklindedir. Burada ise baştaki f dosyanın türünün normal dosya olduğunu gösterir. Yani iki bozulma arasındaki fark dosya türlerinin farklı olmasıdır. Cevabımızın doğru olduğunu kontrol etmemiz için –c bayrağı ile çalıştıralım.



-c bayrağını kullandığımızda inode 10’nin yetim olduğunu yani tahsis edilmediğini söylüyor. Burdan yola çıktığımızda cevabımız doğrudur. –s bayrakları ile farklı dosya sistemlerinde aynı bozulmayı görelim.

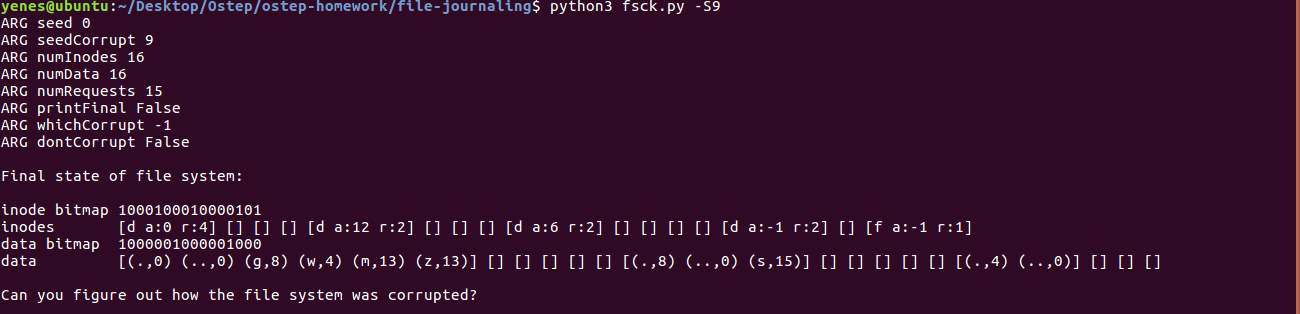
Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde inode bitmap’inde tahsis edilmemesine rağmen 10. İnode doludur.

Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde inode bitmap’inde tahsis edilmemesine rağmen 8. İnode doludur.



Son olarak Dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde de inode bitmap’inde tahsis edilmemesine rağmen 11. İnode doludur. Onarım aracı böyle bir durumla karşılaştığında tahsis edilmeyen bölgeyi siler ve böylece bozulma ortadan kalkar.

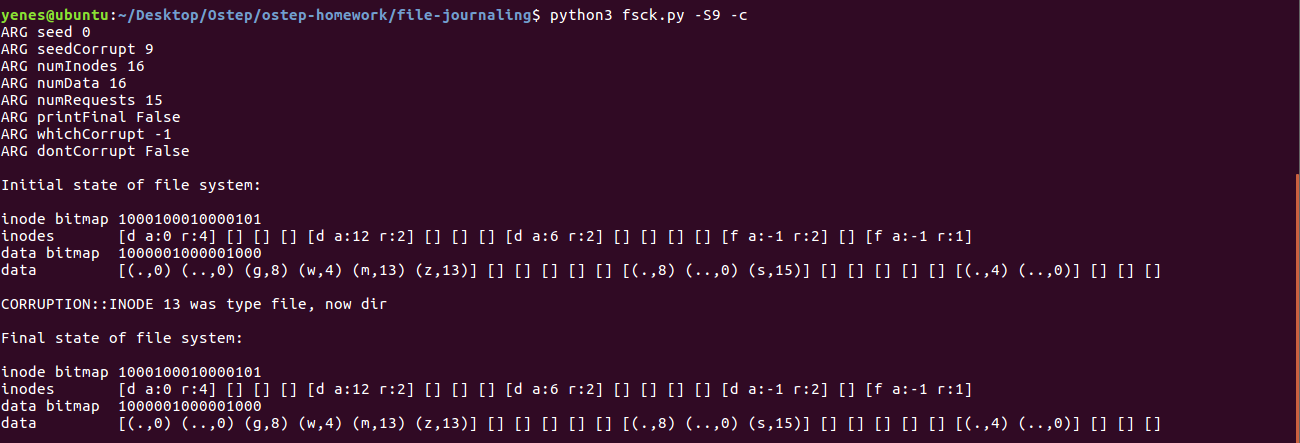
1. Tohumu -S 9 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Cevabınızı kontrol etmek için -c kullanın. Bu durumda bir kontrol ve onarım aracı hangi bilgi parçasına güvenmelidir?



Tohumu –S9 bayrağı ile değiştirdiğimizde tohumu 0 olan dosya sistemindeki bozulmayı görüyoruz. Bozulmayı tespit etmek için –D bayrağını kullanarak bozulmaya uğramamış dosya sistemiyle karşılaştıralım.

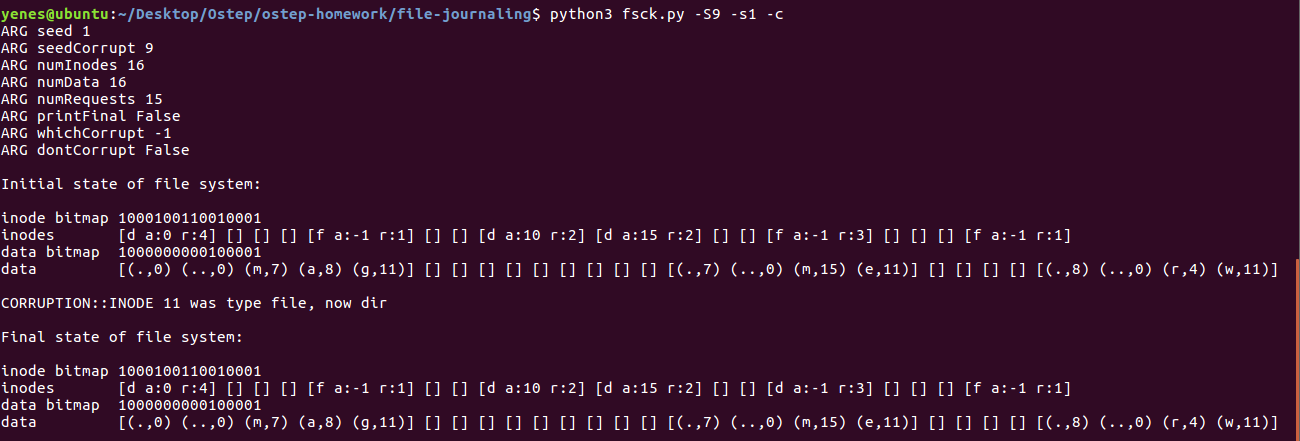


Karşılaştırdığımızda 13. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir. Burada dikkatli olursak a:-1 yani dosyamız boştur. Cevabın doğrulunu kontrol etmek için –c bayrağını çalıştıralım.

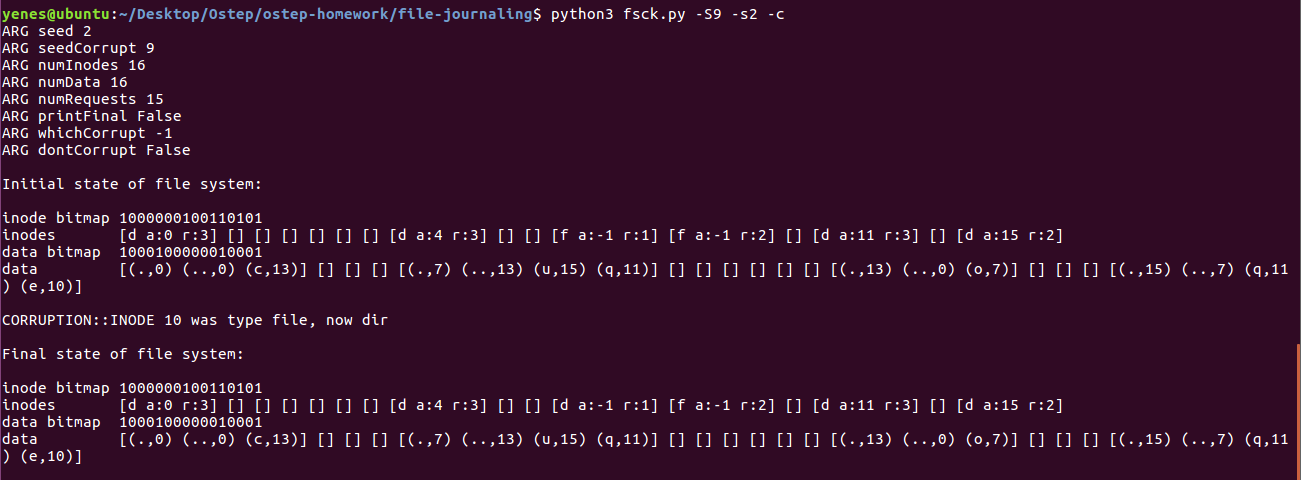


-c bayrağını çalıştırdığımızda inode 13’ün tipinin normal dosya türü olduğunu fakat şuan dizin türünde olduğunu söylüyor ve böylelikle cevabımızın doğru olduğunu görüyoruz.

-s bayrağı ile farklı dosya sistemlerinde bozulmaya bakalım.



Dosya sistemi tohumunu –s1 bayrağı ile çalıştırdığımızda 11. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir.

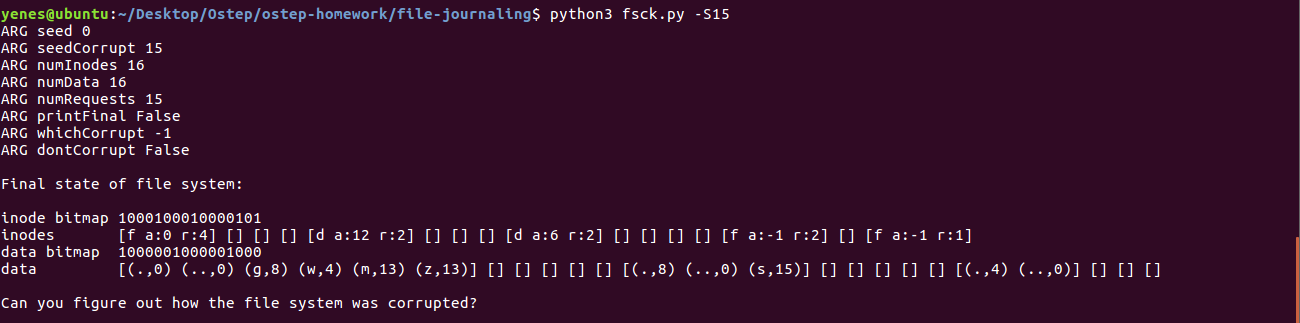


Dosya sistemi tohumunu –s1 bayrağı ile çalıştırdığımızda 10. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir.

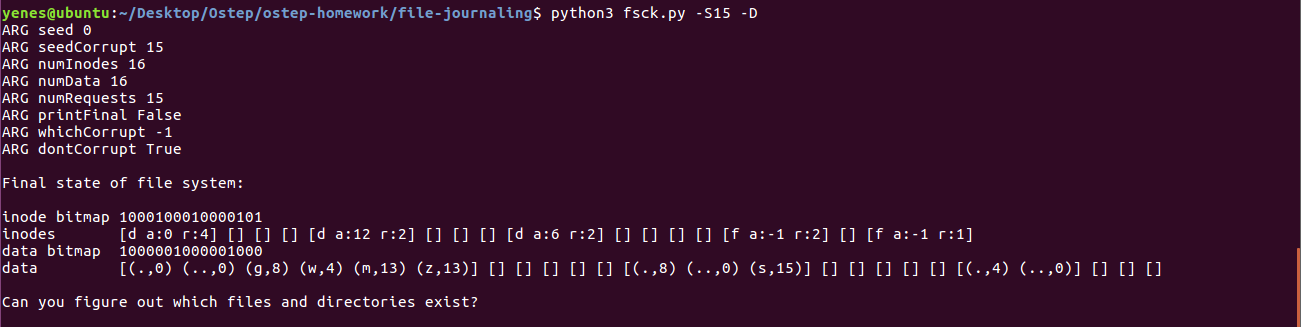


Son olarak Dosya sistemi tohumunu –s3 bayrağı ile çalıştırdığımızda 15. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir. Onarım aracı tahsis edilen her inode’un geçerli bir tür alanına sahip olduğundan emin olur eğer inode alanlarıyla ilgili kolayca düzeltilemeyen sorunlar varsa, inode şüpheli olarak kabul edilir ve onarım aracı tarafından temizlenir ve inode bitmap uygun olarak güncellenir

1. Tohumu -S 15 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Cevabınızı kontrol etmek için -c kullanın. Bu durumda bir onarım aracı ne yapabilir? Onarım mümkün değilse, ne kadar veri kaybedilir?



Tohumu –S15 bayrağı ile değiştirdiğimizde tohumu 0 olan dosya sisteminin bozulmasını görüyoruz. Bozulmanın nedenini –D bayrağı ile bozulma olmayan dosya sistemini çalıştırıp karşılaştırdığımızda bulacağız.



Karşılaştırdığımızda 0. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir fakat önceki sorudan farkı önceki dosya boş iken bu dosyamız doludur ve işaret eder. Cevabın doğrulunu kontrol etmek için –c bayrağını çalıştıralım.

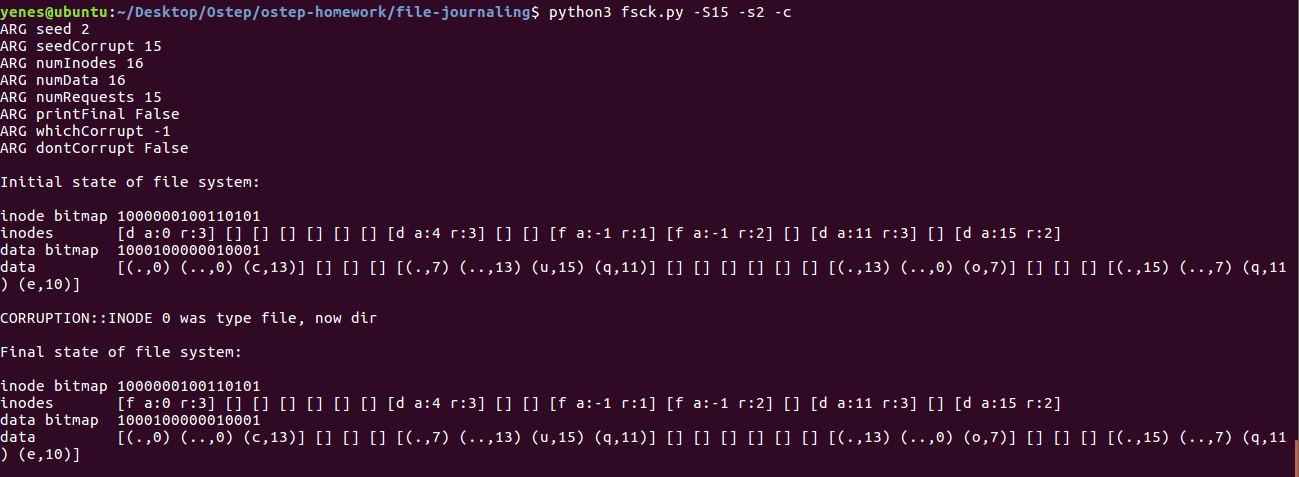


-c bayrağını çalıştırdığımızda inode 0’ın tipinin normal dosya türü olduğunu fakat şuan dizin türünde olduğunu söylüyor ve böylelikle cevabımızın doğru olduğunu görüyoruz.

-s bayrağı ile farklı dosya sistemlerinde bozulmaya bakalım.



Dosya sistemi tohumunu –s1 bayrağı ile çalıştırdığımızda 0. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir.

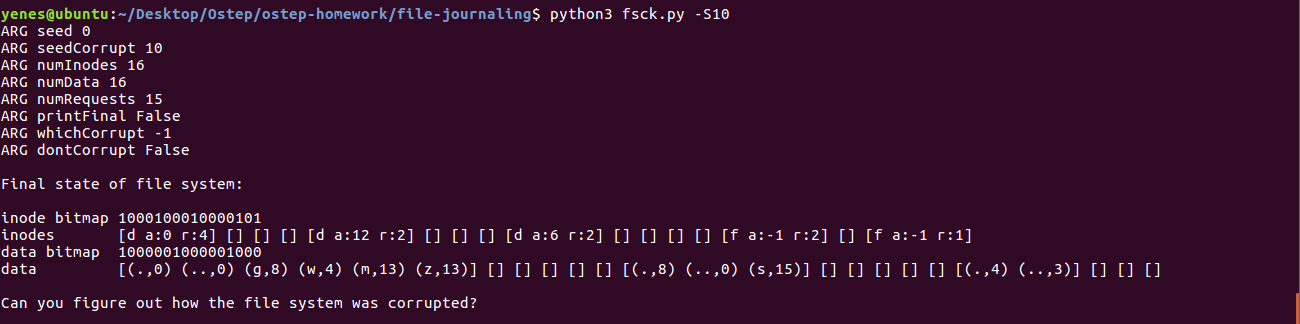


Dosya sistemi tohumunu –s2 bayrağı ile çalıştırdığımızda 0. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir.

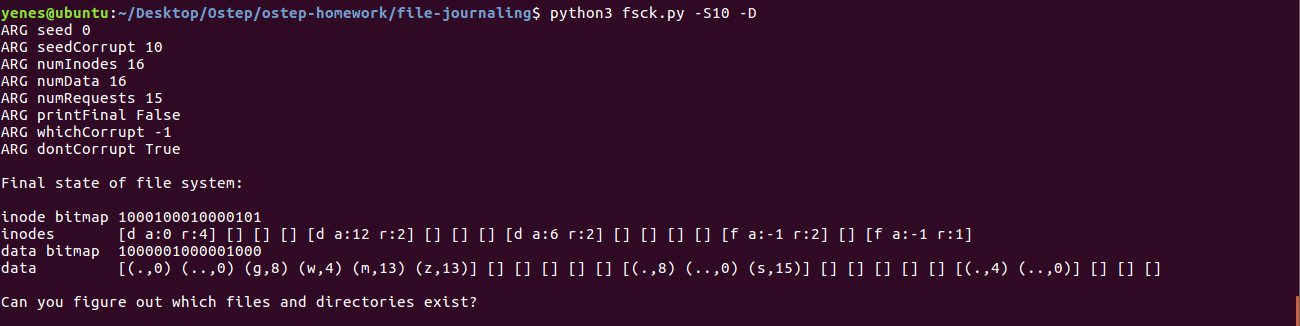


Son olarak Dosya sistemi tohumunu –s3 bayrağı ile çalıştırdığımızda 0. İnode dosya türü normal dosya iken bozulmaya uğradığında tür dizin olarak değişmiştir. Cevapların aynıdır fakat dosya sistemleri gördüğümüz üzere farklıdır. Onarım aracı İnode Stati kontrol ederek bozulmayı onarabilir fakat dizin içeriği normal dosya sistemine dönüştüğü kadar veri kaybı yaşar.

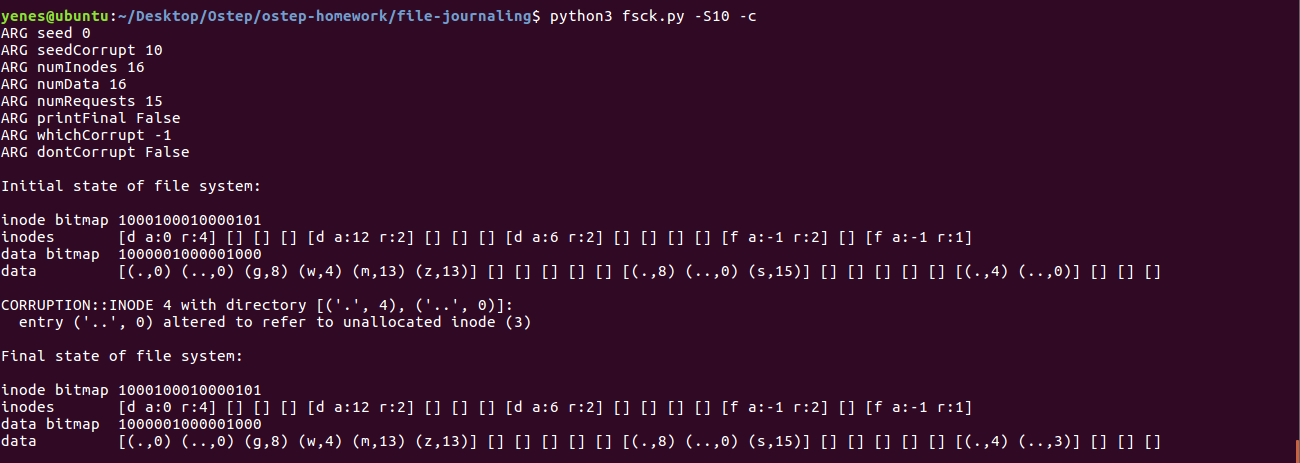
1. Tohumu -S 10 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Cevabınızı kontrol etmek için -c kullanın. Burada dosya sistemi yapısında onarıma yardımcı olabilecek bir fazlalık var mı?



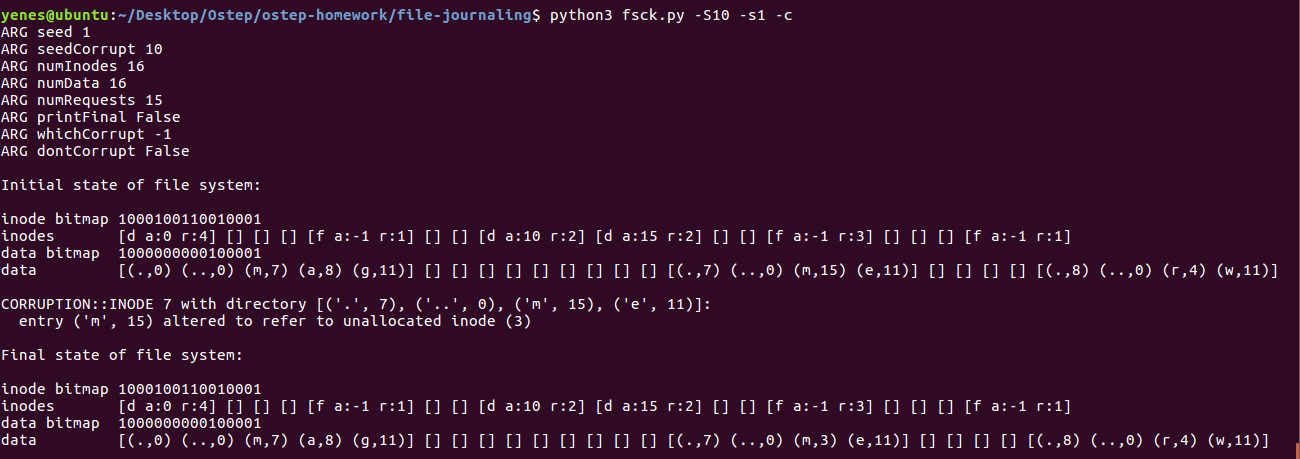
Tohumu –S15 bayrağı ile değiştirdiğimizde tohumu 0 olan dosya sisteminin bozulmasını görüyoruz. Bozulmayı anlamamız için –D bayrağı ile bozulma meydana gelmemiş haliyle karşılaştıralım.



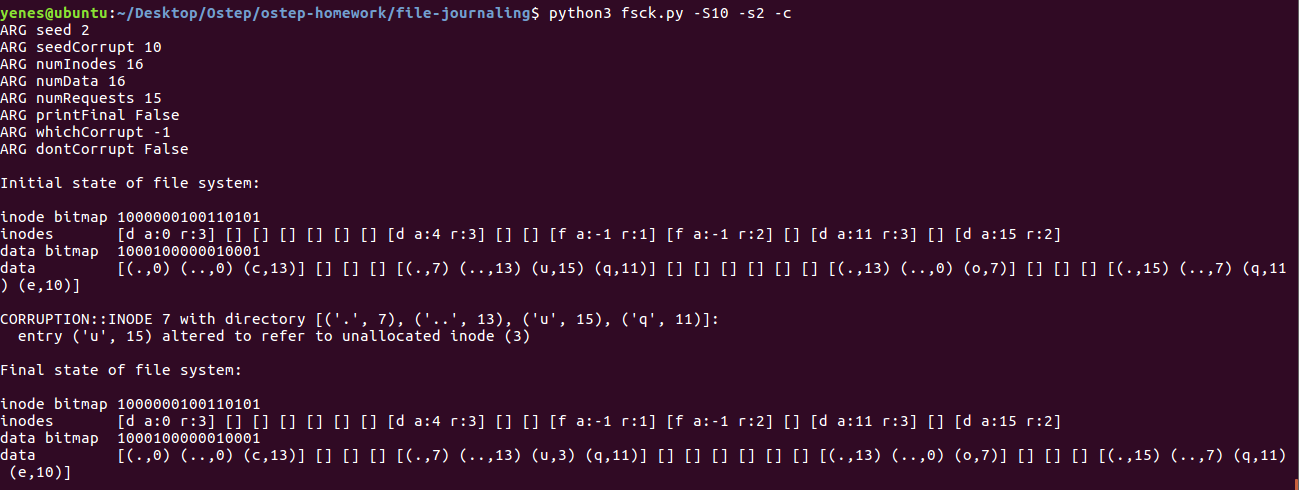
Karşılaştırdığımızda 4. İnode’ın işaret ettiği data bölgesindeki 12. adres bloğundaki (..,0) verisi (..,3) olarak değiştiğini ve bozulma meydana geldiğini görüyoruz. Cevabımızı –c bayrağını çalıştırarak kontrol edelim.



-c bayrağını çalıştırdığımızda giriş (‘..’, 0) ayrılmamış inode’a (3) referans için değiştirildiğini söylüyor yani cevabımız doğrudur. –s ile dosya sistemi tohumlarımızı değiştirip farklı dosya sistemlerinde bozulmamıza tekrar bakalım.



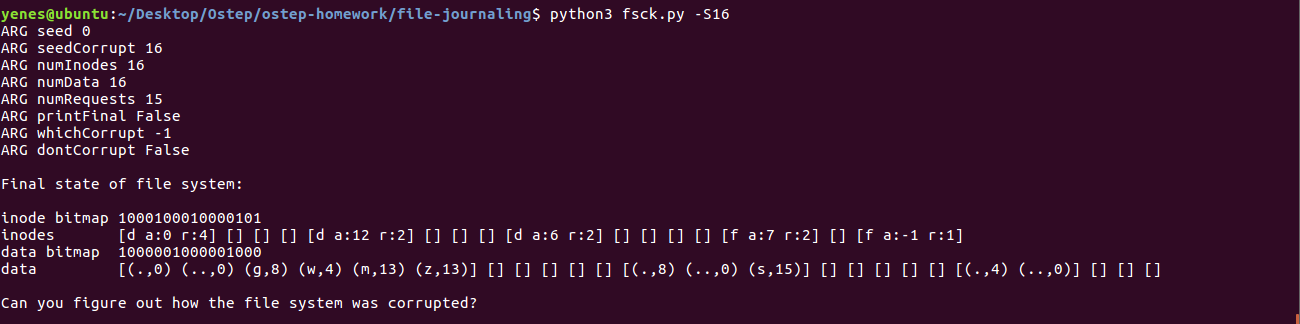
Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde 7. İnode’un işaret ettiği data bölgesindeki 10. adres bloğundaki (m,15) verisi (m,3) olarak değiştiğini görüyoruz.



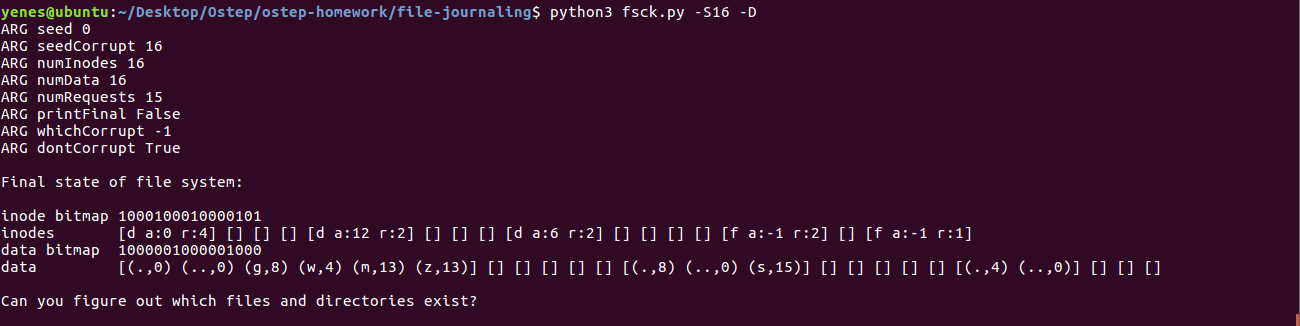
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde 7. İnode’un işaret ettiği data bölgesindeki 4. adres bloğundaki (u,15) verisi (u,3) olarak değiştiğini görüyoruz.



Son olarak Dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde 2. İnode’un işaret ettiği data bölgesindeki 13. adres bloğundaki (..,0) verisi (..,13) olarak değiştiğini görüyoruz.

1. Tohumu -S 16 ve -S 20 olarak değiştirin; hangi tutarsızlığı görüyorsunuz? Cevabınızı kontrol etmek için -c kullanın. Onarım aracı sorunu nasıl çözmelidir?

-S16 bayrağı ile tohumu 0 olan dosya sistemimizin yeni bir bozulma durumunu görüyoruz. Bozulmayı daha iyi anlamamız için –D bayrağı ile bozulmaya uğramamış halini karşılaştırmamız gerekiyor.



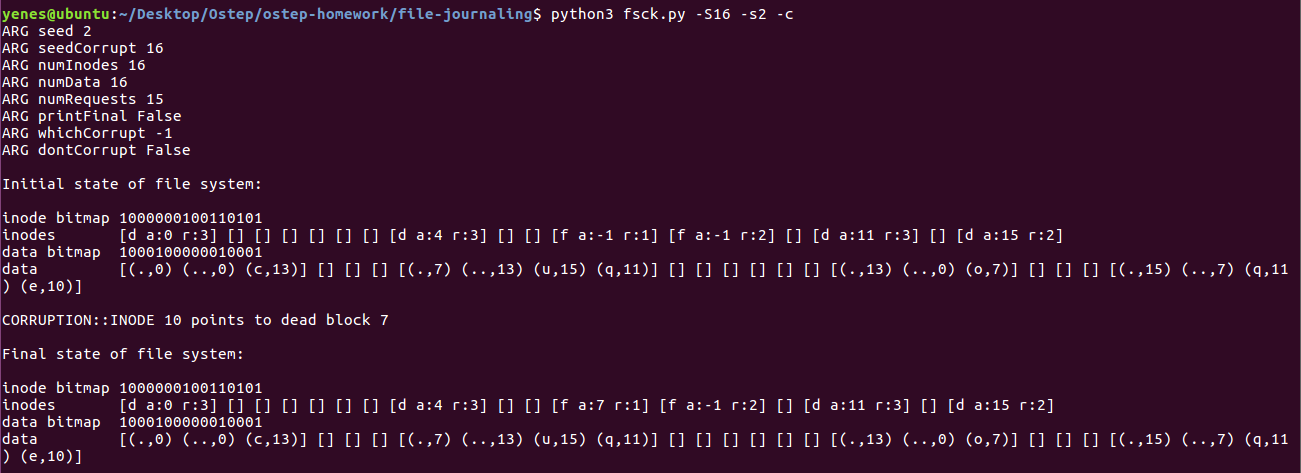
Karşılaştırdığımızda 13. İnode’daki normal dosya türümüz normalde -1 ile boşken bozulmaya uğradığında 7. Bölgeyi işaret etmiş gibi gösteriyor yani ölü bir bloğa işaret etmiş oluyor. –c bayrağını çalıştırarak cevabımızı kontrol edelim.



-c bloğunu çalıştırdığımızda inode 13’ün ölü blok 7yi işaret ettiğini söylüyor böylelikle cevabımızın doğru olduğunu görüyoruz. Diğer dosya sistemlerinde hatamıza bakmak için –s bayraklarını kullanalım.



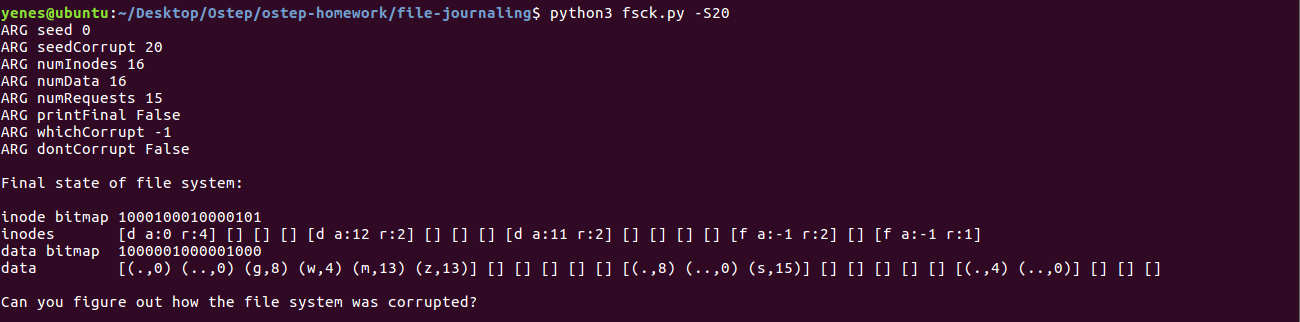
Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde inode 11’in ölü blok 6’yı işaret ettiğini görüyoruz.



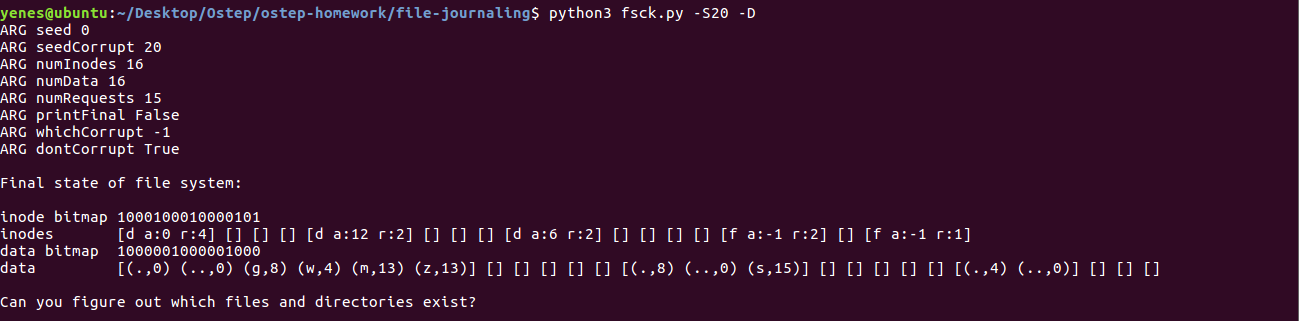
Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde inode 10’un ölü blok 7’yi işaret ettiğini görüyoruz.



Son olarak dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde inode 15’in ölü blok 6’yı işaret ettiğini görüyoruz. Bir de –S20 bayrağı ile çalıştırmayı deneyelim.



-S20 bayrağı ile tohumu 0 olan dosya sisteminde bozulmayı çalıştırıyoruz. Bozulmayı tespit etmek için –D bayrağını çalıştırarak bozulmaya uğramamış haliyle karşılaştırmamız gerekiyor.



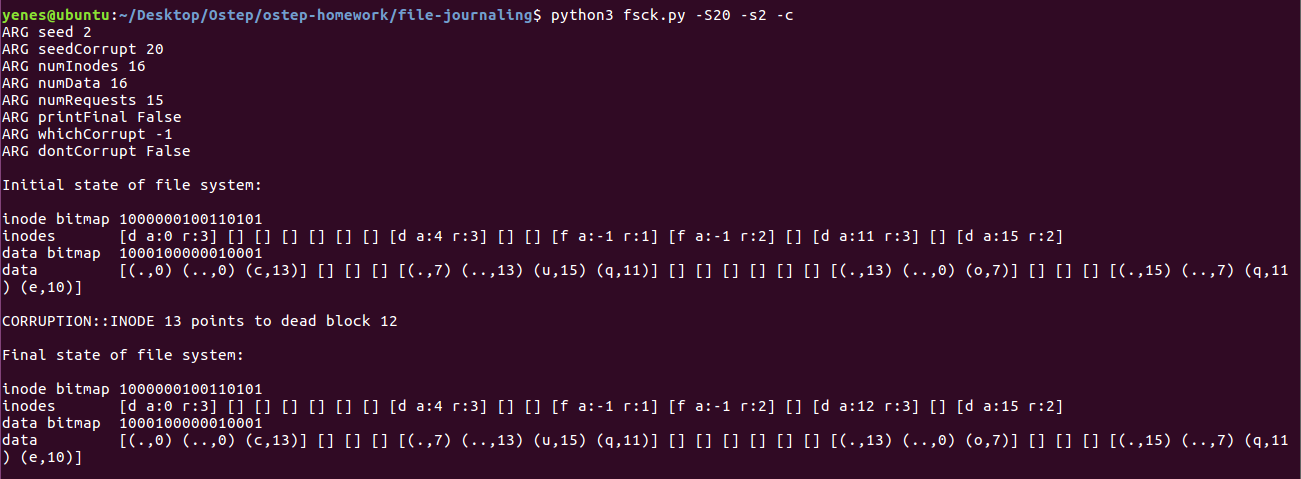
Karşılaştırdığımızda türü dizin olan 8. İnode’umuz normalde 6. Adresi işaret ederken bozulmaya uğradığında 11. Adresi işaret ediyor yani ölü bir bloğa işaret ediyor. Cevabımızı kontrol etmek için –c bayrağını kullanalım.



-c bloğunu çalıştırdığımızda inode 8’ün ölü blok 11’i işaret ettiğini söylüyor böylelikle cevabımızın doğru olduğunu görüyoruz. Diğer dosya sistemlerinde hatamıza bakmak için –s bayraklarını kullanalım.



Dosya sistemi tohumunu –s1 bayrağı ile değiştirdiğimizde inode 8’in ölü blok 11’yı işaret ettiğini görüyoruz.



Dosya sistemi tohumunu –s2 bayrağı ile değiştirdiğimizde inode 13’ün ölü blok 12’yı işaret ettiğini görüyoruz.



Son olarak dosya sistemi tohumunu –s3 bayrağı ile değiştirdiğimizde inode 3’ün ölü blok 11’i işaret ettiğini görüyoruz. Onarım aracı inode’u kontrol edip bozulmayı bulmalı ve düzeltmelidir.